فصلنامه علمی- پژوهشی اخلاق پژوهی
سال ششم • شماره چهارم • زمستان ۱۴۰۲
Quarterly Journal of Moral Studies
Vol. 6, No. 4, Winter 2024

واکاوی چالش‌های پیاده‌سازی هوش اخلاقی در هوش مصنوعی

محدثه قوامی‌پور سرشکه* | امیر رضا محمودی**

10.22034/ethics.2024.51227.1650

چکیده

پیاده‌سازی یک نظام اخلاقی در رفتار سیستم‌های هوش مصنوعی، صرفاً نیازمند هوش انتزاعی نیست، بلکه مستلزم شکل خاصی از هوش انسانی و مصنوعی است. عاملان هوشمند مصنوعی رفتار و ویژگی‌های بارزی مثل منفرد بودن، پیچیده بودن، تخصصی بودن، پایدار و خودگردان بودن دارند. کُد اخلاقی انسان به نحو جبران‌ناپذیری معیوب است: این اخلاق ترجیح، پاسخ‌گویی (نه مسئولیت) و عدم آزادی است که با ابداع سیستم‌های اخلاقی، گردش ارزش‌های آرمانی و آموزش اخلاق نمی‌توان آن را جبران کرد، اما چنین اخلاق معیوبی برای نهادینه سازی، نیازمند ابزارهای کاملی است که نه‌تنها هوش، بلکه وجدان و حتی روح را به ماشین می‌بخشد. در این مقاله، پس از بیان مسئله و مفهوم‌شناسی هوش اخلاقی و اشاره به ویژگی‌های عاملان هوشمند مصنوعی، به دشواری‌های اعمال قواعد اخلاقی برای عوامل هوشمند و همچنین موانع پیشِ روی هوش مصنوعی اخلاقی اشاره شده و مهم‌ترین چالش‌های پیاده‌سازی هوش اخلاقی در هوش مصنوعی، واکاوی و تحلیل شده است.

کلیدواژه‌ها

هوش مصنوعی، هوش اخلاقی، اخلاق ماشینی، کُد اخلاقی.

* دانشجوی دکتری حقوق کیفری و جرم‌شناسی، گروه حقوق دانشگاه آزاد اسلامی لاهیجان، لاهیجان، ایران.
(نویسنده مسئول) | mohadesehghavmipour@liau.ac.ir

** استادیار گروه حقوق، دانشگاه آزاد اسلامی واحد لاهیجان، لاهیجان، ایران. | Amirreza.mahmodi@liau.ac.ir

تاریخ پذیرش: ۱۴۰۳/۰۱/۰۸ ▣ تاریخ تأیید: ۱۴۰۳/۰۳/۰۵

■ قوامی‌پور سرشکه، محدثه؛ محمودی، امیر رضا. (۱۴۰۲). واکاوی چالش‌های پیاده‌سازی هوش اخلاقی در هوش مصنوعی. فصلنامه اخلاق‌پژوهی. ۶(۲۱)، ۲۳-۴۸. doi: 10.22034/ethics.2024.51227.1650

مقدمه

در دنیای امروز که تعامل انسان و ماشین به شکلی بی‌سابقه گسترش یافته، پیدایش محیط‌های فنی مصنوعی و تبدیل ماشین‌های اطلاعاتی به ماشین‌های هوشمند، گذار عملی به یک محیط فکری و فنی را رقم زده که منجر به ظهور حوزه جدید تحقیقاتی اخلاق ماشینی شده است (کریمی واقف و عبدخدایی، ۱۴۰۰، ص ۷۱). تا کنون تأثیرات اخلاقی عملکرد ماشین‌ها نادیده گرفته شده و مسئولیت پیامدهای آن به کاربر، طراح یا مالک آنها محول شده است (بلبلی قادیکلای و پارسانیا، ۱۴۰۲، ص ۷۸۱).

اهمیت و ضرورت بررسی و پرداختن به این موضوع از آنجا ناشی می‌شود که موجودات انسانی و مصنوعی اکنون در اشکال جدیدی در حرکت و تعامل هستند. ماشین‌ها کارکردهای متعدد و معناداری مرتبط با انسان و جامعه را انباشته کرده‌اند. عاملان مصنوعی نه‌تنها برای یاری رساندن به انسان‌ها، بلکه برای جایگزینی آنها در فرایندهای مختلفی همچون تولید، کار، خدمات، ارتباطات، تحقیق، آموزش و سرگرمی خلق شده‌اند. بنابراین، رفتار این عاملان دارای اهمیت اخلاقی است.

این دنیای رو به مصنوعی شدن که از تعامل انسان-ماشین پدید آمده، نه‌تنها پیچیدگی فزایندهٔ ماشین را موجب می‌شود، بلکه پیچیدگی فزایندهٔ خود بشریت و ارزش‌های اخلاقی آن را نیز در پی دارد. گونه انسانی در همه ابعاد زیستی، روانی، اجتماعی و فرهنگی در حال تکامل است و اینک به سمت مصنوعی شدن گام برمی‌دارد (Gregory, 2000, pp. 171-179; Hillis, 2001, p. 59).

مسئله اصلی این پژوهش، چالش‌های پیاده‌سازی هوش اخلاقی در سیستم‌های هوش مصنوعی است. هوش اخلاقی به توانایی یک عامل هوشمند در درک مفاهیم اخلاقی و تصمیم‌گیری بر اساس آنها اشاره دارد. با توجه به گسترش روزافزون کاربردهای هوش مصنوعی در عرصه‌های مختلف زندگی بشر، تعبیه این ویژگی در عاملان مصنوعی اهمیت بسیار دارد.

پرسش اصلی این پژوهش آن است که چالش‌ها و موانع پیشِ روی پیاده‌سازی هوش اخلاقی در هوش مصنوعی کدامند و چگونه می‌توان آنها را شناسایی و تحلیل کرد؟ از طریق مطالعه کتب، مقالات و دیدگاه‌های صاحب‌نظران حوزهٔ اخلاق ماشینی و هوش مصنوعی، این پژوهش به دنبال آن است که چالش‌های کلیدی این موضوع را شناسایی، تبیین و تحلیل کند.

به باور نگارنده، شناخت این موانع می‌تواند گامی مهم در جهت یافتن راهکارهایی مناسب برای غلبه بر آنها و در نتیجه پیاده‌سازی موفق هوش اخلاقی در عاملان هوشمند باشد. این امر با

توجه به گسترش روزافزون کاربردهای هوش مصنوعی در ابعاد گوناگون زندگی انسانی، ضرورتی انکارناپذیر است.

پیشینهٔ پژوهش

بلبلی قادیکلایی و پارسانیا (۱۴۰۲) در پژوهشی با عنوان «مروری نظام‌مند بر دلالت‌های اخلاقی استفاده از هوش مصنوعی در فناوری‌های دیجیتال و نسبت آن با اخلاق شکوفایی» به بررسی زمینه اخلاق هوش مصنوعی در فناوری‌های دیجیتال پرداخته‌اند. نقطه قوّت اصلی آن، استفاده از روش مرور نظام‌مند ادبیات و فراترکیب کیفی است که به پژوهش جامعیت و روایی می‌بخشد. شناسایی مفاهیم کلیدی اخلاقی در هوش مصنوعی از طریق بررسی مقالات مختلف، ارتباط دادن این مفاهیم با الگوها و ساختارهای فناوری‌های دیجیتال، بررسی نسبت آنها با نظریه اخلاق شکوفایی و ارائه یک نقشهٔ مفهومی برای طبقه‌بندی و تفسیر یافته‌ها، از دیگر نکات قوّت این مطالعه به شمار می‌رود. با این حال، محدود بودن به دو پایگاه داده خاص، تمرکز صرف بر فناوری‌های دیجیتال، عدم بررسی عمیق چالش‌های پیاده‌سازی عملی و محدودیت به مقایسه با تنها یک نظریهٔ اخلاقی، از نقاط ضعف این پژوهش است. در مجموع، این مطالعه تلاشی ارزشمند، اما نیازمند پژوهش‌های بیشتر و بررسی‌های میدانی برای کاربرد عملی است. پژوهش حاضر با تمرکز مستقیم بر چالش‌ها و موانع پیاده‌سازی عملی هوش اخلاقی در سیستم‌های هوشمند، رویکردی کاربردی‌تر و نزدیک‌تر به مسائل واقعی و عملیاتی در این حوزه را در پیش می‌گیرد. بدین ترتیب، ظرفیت بیشتری برای ارائهٔ راهکارها و پیشنهادهای عملی جهت غلبه بر چالش‌های شناسایی شده خواهد داشت. در نهایت، این مقاله می‌تواند به ارائهٔ یک چارچوب عملیاتی برای غلبه بر موانع پیاده‌سازی هوش اخلاقی در عاملان هوشمند واقعی بپردازد که مکمّل مناسبی برای پژوهش‌های نظری و مفهومی پیشین خواهد بود.

کریمی واقف و عبدخدایی (۱۴۰۰) در پژوهشی با عنوان «چالش‌های فرا روی کاربست اخلاق در ماشین‌های هوشمند؛ با تمرکز بر رویکرد اصل‌گرایی در اخلاق» به چالش‌های پیاده‌سازی اصول اخلاقی بر مبنای رویکرد بنیادگرایی اخلاقی در سیستم‌های هوش مصنوعی با استفاده از روش اجرای از بالا به پایین (روشی که در آن اصول و قواعد کلی اخلاقی ابتدا تعیین و سپس در سطوح پایین‌تر به صورت جزئی‌تر پیاده می‌شوند) می‌پردازد. نقاط قوّت آن شامل

بررسی یکی از رویکردهای مهم فلسفهٔ اخلاق، تحلیل و نقد ابعاد مختلف این مسئله، ارائهٔ تعاریف و مفاهیم کلیدی و بررسی انتقادات و دیدگاه‌های دیگر متفکران در این زمینه است. با این حال، محدود بودن به یک رویکرد و روش خاص، عدم ارائه راهکارهای عملی و محدود شدن به بررسی نظری و فلسفی از نقاط ضعف احتمالی این مقاله است. در مجموع، این مقاله می‌تواند بینش ارزشمندی در خصوص چالش‌های مورد نظر ارائه دهد، اما برای کاربرد عملی نیازمند پژوهش‌ها و بررسی‌های میدانی بیشتری خواهد بود. این مقاله با تمرکز مستقیم بر چالش‌های پیاده‌سازی عملی هوش اخلاقی در سیستم‌های هوشمند، بررسی ویژگی‌های خاص این عاملان، پرداختن به مفاهیم کلیدی نظیر وجدان و ساختار اخلاقی و ارائهٔ راهکارهایی برای کشف و اجرای هوش اخلاقی، رویکردی کاربردی‌تر را در پیش گرفته است. افزون بر این، با معرفی چارچوبی برای هنجارهای اخلاقی مناسب برای عاملان انسانی و مصنوعی، گام‌های عملی‌تری برای تحقّق هوش اخلاقی برداشته که مکمّل خوبی برای مطالعات نظری و مفهومی پیشین در این حوزه خواهد بود.

فلوریدی و کاولز (۲۰۲۳) در مقاله‌ای با عنوان «اخلاق در هوش مصنوعی: گرایش به یک شاخه تفکیک شده و هدفمند» استدلال می‌کنند که اخلاق هوش مصنوعی باید از مطالعات فلسفی کلاسیک در مورد اخلاق رها شود و روی مسائل عملی و کاربردی که صنعت هوش مصنوعی با آن روبه رو است، متمرکز شود. آنها معتقدند که اخلاق در زمینهٔ هوش مصنوعی باید یک رشته تخصصی و کاربردی شود که با چالش‌های واقعی هوش مصنوعی سر و کار دارد. این نویسندگان پیشنهاد می‌کنند که اخلاق هوش مصنوعی باید از یک رویکرد توصیف‌گرایانه و انتزاعی به سمت یک رویکرد تجویزی و عملیاتی حرکت کند. به جای بحث‌های فلسفی کلی دربارهٔ ارزش‌ها و اصول اخلاقی، باید بر راه حل‌های عملی برای مشکلات مشخص هوش مصنوعی از جمله عدم تعصب، شفافیت و مسئولیت‌پذیری تمرکز شود . آنها همچنین پیشنهاد می‌کنند که اخلاق هوش مصنوعی باید به صورت میان‌رشته‌ای و با همکاری تنگاتنگ فیلسوفان، متخصصان هوش مصنوعی، قانونگذاران و دیگر ذینفعان پیش برود. پژوهش حاضر با رویکرد سیستمی و ساختاری به هوش مصنوعی و اخلاق، در نظر گرفتن چالش‌های اعمال اخلاق برای هر دو عامل انسانی و مصنوعی، تمرکز بر ایجاد یک چارچوب و سیستم اخلاقی برای هوش مصنوعی و هدف گرفتن گام‌های عملی به سمت پیاده‌سازی قواعد اخلاقی در هوش مصنوعی، نوآوری‌های مهمی نسبت به مقالهٔ قبلی دارد که بیشتر جنبه‌های نظری را مورد بحث قرار داده بود. این مقاله می‌کوشد با

دیدگاهی جامع‌تر، چالش‌های عملی اعمال اخلاق در هوش مصنوعی را بررسی کند.

روش‌شناسی

این پژوهش از یک رویکرد تحلیلی-تفسیری برای بررسی «دشواری‌های مفهومی و فنی اعمال قواعد اخلاقی برای عوامل انسانی و دارای هوش مصنوعی» و ارائه یک «ساختار سیستم اخلاقی» جهت «کشف هوش اخلاقی، گامی به سوی اجرای قواعد اخلاقی» استفاده می‌کند. ابتدا مفهوم «هوش به عنوان یک سیستم با ساختار داخلی» از طریق مطالعه تحلیلی-تفسیری جامع ادبیات در حوزه‌های هوش مصنوعی، روان‌شناسی، اخلاق و سیستم‌های پیچیده مورد واکاوی قرار می‌گیرد. سپس چندین سیستم یا عامل هوش مصنوعی که شامل «ویژگی‌های عوامل «دارای هوش مصنوعی» است، به عنوان موارد مطالعاتی انتخاب می‌شوند. مستندات، گزارش‌ها، اسناد و کدهای منتشر شده مرتبط با این سیستم‌ها به صورت تحلیلی-تفسیری بررسی می‌شوند تا دشواری‌های مفهومی و فنی اعمال قواعد اخلاقی" در آنها کشف و تفسیر شود.

مفاهیم «وجدان اخلاقی» و «هنجارهای اخلاقی» نیز از طریق مطالعه تفسیری پیشینه نظری و تحلیل تفسیری موارد مطالعاتی بازکاوی می‌شوند. یافته‌های حاصل از این فرایندهای تحلیلی-تفسیری، در یک تحلیل تفسیری نهایی تلفیق شده و منجر به شکل‌گیری یک «ساختار سیستم اخلاقی» جامع برای ادغام اخلاق در عوامل هوشمند انسانی و مصنوعی می‌گردد که «گامی به سوی اجرای قواعد اخلاقی» و «کشف هوش اخلاقی» در این عوامل به‌حساب می‌آید.

۱. هوش اخلاقی و هوش مصنوعی

هوش اخلاقی یا هوشمندی اخلاقی، یک مفهوم چند بُعدی است که شامل توانایی تشخیص، استدلال و عمل بر اساس اصول و ارزش‌های اخلاقی در موقعیت‌های مختلف می‌شود. این مفهوم، جنبه‌ای حیاتی در تصمیم‌گیری، به‌ویژه در شرایط پیچیده و مبهم که در آن ملاحظات اخلاقی نقش دارند، به‌حساب می‌آید. هوش اخلاقی دارای چندین جزء کلیدی است:

نخست، حساسیت اخلاقی که به توانایی شناسایی و تفسیر مسائل و معضلات اخلاقی در یک موقعیت خاص اشاره دارد (Narvaez & Endicott, 2009, p 156). دوم، قضاوت اخلاقی که به ظرفیت ارزیابی مُجاز یا نامجاز بودن اخلاقی روش‌های مختلف عمل اشاره دارد. سوم، انگیزش اخلاقی

که به میل درونی برای اولویت‌بندی ارزش‌های اخلاقی و عمل بر اساس ملاحظات اخلاقی اشاره دارد و در نهایت، رفتار اخلاقی که به توانایی واقعی برای پیاده‌سازی تصمیمات و اقدامات اخلاقی، حتی در برابر موانع یا فشارهای رقابتی اشاره می‌کند (Narvaez & Lapsley, 2005, p. 141).

با پیشرفت فناوری و پیچیده‌تر شدن سیستم‌های هوش مصنوعی و ادغام آنها در حوزه‌های مختلف، تجهیز آنها به هوش اخلاقی برای اطمینان از هماهنگی با ارزش‌های انسانی و به حداقل رساندن خطرات و عواقب نامطلوب بالقوه، امری حیاتی است (Wallach & Allen, 2009, p 63). پیاده‌سازی هوش اخلاقی در سیستم‌های هوش مصنوعی یک چالش پیچیده است که نیازمند تلاش‌های میان‌رشته‌ای از حوزه‌هایی همچون فلسفه، روان‌شناسی، علوم رایانه و علوم شناختی است (Cervantes et. al., 2020, p. 2).

هوش مصنوعی به توانایی ماشین‌ها، کامپیوترها یا سیستم‌های رایانه‌ای برای شبیه‌سازی رفتارها و قابلیت‌های هوش انسانی مانند یادگیری، استدلال، حل مسئله، برنامه‌ریزی و تصمیم‌گیری اشاره دارد (Russell & Norvig, 2020, p. 1). این حوزه شامل دو رویکرد اصلی است: نخست، هوش مصنوعی قوی یا هوشمند که به معنای توانایی یک سیستم برای انجام هر کاری است که یک انسان می‌تواند انجام دهد (Goertzel & Pennachin, 2007, p. 7). دوم، هوش مصنوعی ضعیف یا هوشمند محدود که به توانایی یک سیستم برای انجام یک کار خاص به شکل بهتر از انسان، اما در یک زمینه محدود اشاره دارد (Russell & Norvig, 2020, p. 7). هوش مصنوعی در حوزه‌های مختلفی مانند پردازش زبان طبیعی، بینایی کامپیوتری، سیستم‌های توصیه‌گر، هوش مصنوعی تقویتی و یادگیری ماشین کاربرد دارد (Goodfellow et al., 2016). این فناوری پیشرفته امکانات جدیدی را در زمینه‌های مختلف از جمله پزشکی، مهندسی، تجارت و امنیت ایجاد کرده است، اما همچنین چالش‌ها و نگرانی‌هایی را در مورد مسائل اخلاقی، حریم خصوصی و کنترل انسانی بر این سیستم‌ها به همراه داشته است.

۲. اخلاق جدید برای عاملان انسانی-مصنوعی

اخلاق پردازش اطلاعات برای همهٔ کسانی که در محیط فنی-فکری جدید از رایانه استفاده می‌کنند، مفید است. این یک اخلاق حرفه‌ای نیست، بلکه برای کارکنان رایانه و شبکه در مشاغل مختلفی که اطلاعات را به این شکل پردازش و انتقال می‌دهند، در نظر گرفته شده است. موضوعاتی مثل

حفاظت از مالکیت نرم‌افزار، تأمین هویت و حریم خصوصی کاربران و حتی اشتراک‌گذاری و حفاظت از آئین‌نامه اینترنتی برای این حوزه اخلاقی به عنوان مشخصه پذیرفته شده‌اند.

اخلاق محاسباتی، از رایانه‌ها در حوزهٔ اخلاق فلسفه برای حلِّ مسائل اخلاقی نظری و عملی استفاده می‌کند؛ ابزارها و روش‌های آموزش و یادگیری مبتنی بر رایانه پذیرفته و توسعه داده می شوند. این حوزهٔ اخلاقی از نتایج تحقیقات فناوری اطلاعات پشتیبانی شده با تکنیک‌های هوش مصنوعی مبتنی بر دانش و همچنین مزیّت‌های یادگیری تله‌ماتیک،[1] (دورا داده‌ورزی) توزیع‌شده، انعطاف‌پذیر و چندرسانه‌ای استفاده می‌کند. نظریه‌های اخلاقی معتبر همچنین با روش‌های محاسباتی که تجزیه و تحلیل اساسی تصمیم را در مسائل اخلاقی دشوار امکان‌پذیر می‌سازند، بررسی می‌شوند (محرابی، خراشادی زاده، ۱۴۰۲، ص۳۵۵) با پیشرفت مُدلسازی و شبیه‌سازی، پیامدهای اخلاقی تصمیمات اجتماعی معنادار پیش‌بینی می‌شود.

دانیلسون در سال ۱۹۸۸ با تحقیقاتی که صورت داد، آشکار ساخت که بخش‌های مهمی از اخلاق همیشه مصنوعی بوده است و اذعان کرد که با استفاده از رایانه در این حوزه، ما تنها ویژگی مصنوعی اخلاق را گسترش می‌دهیم (Danielson, 1998, p. 292).

اخلاق ماشینی به خود رایانه مربوط می‌شود؛ ماشین هوشمند ـ همانند انسان ـ تغییراتی در جهان ایجاد می‌کند. تمام فعالیت‌های انسانی دارای اهمیت اخلاقی هستند. یک ماشین با قابلیت‌های مشابه نیازمند کارکردهای اخلاقی است. اخلاق ماشینی حوزهٔ تحقیقاتی جدیدی است که در آن متخصصان کلی فلسفه و متخصصان هوش مصنوعی همکاری می‌کنند؛ (Pana, 2002, p. 69) نتیجه، یک اخلاق مصنوعی خواهد بود که بخشی از فلسفه مصنوعی را تشکیل می‌دهد. اخلاق ماشینی، در واقع، تمایل دارد یک حوزه از تحقیقات هوش مصنوعی باشد، اما بدون یک پایه فلسفی قوی (هستی‌شناسی، ارزش‌شناسی، کاربردشناسی و اخلاق) نمی‌تواند طراحی و ساخته شود. فلسفه کنونی می‌تواند با مسائل خاصی که توسط اخلاق مصنوعی مطرح می‌شود، کنار بیاید؛ زیرا حتی به روش‌های مختلف به سمت فلسفه مصنوعی تکامل می‌یابد (Pana, 2002, p. 69).

اصطلاح «فلسفه مصنوعی» نخستین‌بار توسط فر. لارول[2] استفاده شد و علم تفکر را که با روش‌های ریاضی و فنی توسعه یافته بود، مشخص کرد (Laruelle, 1990, p. 235). پیش از آن دلاتیل[3] در مورد «تفکر سنتتیک» یا مصنوعی کار کرده بود.

1. Telematics
2. Fr. Laruelle
3. P. de Latil

اخلاق جهانی اطلاعات بـا جهـانی شـدن اطلاعـات، ارتباطـات و کـار شـکل گرفتـه اسـت. اخلاق جهانی اطلاعات می‌تواند نه به سادگی مجموع حوزه‌های اخلاقـی جدیـد مـذکور در بـالا، بلکه به طور دقیق‌تر، محصول و سطح جدید فراساختاری از سنتز شکل‌های اخلاقـی فـوق تحـت شرایط اطلاعاتی شدن و فکری شدن تمام فعالیت‌های انسانی در نظر گرفته شـود. جنبـه‌های مهـم اخلاق جهانی اطلاعات که توسط باینوم[1] برجسته شده اسـت، نشـان مـی‌دهـد کـه در ایـن حـوزه اخلاقـی، ارزش‌هـا و رفتارهـای اخلاقـی موضـوعات بحـث و همـاهنگی فراتـر از تفاوت‌هـای جغرافیایی، اجتماعی، سیاسی و حتی فرهنگی می‌شوند (Bynum, 1998, p. 34).

۳. ساختار سیستم اخلاقی

سیستم‌های اخلاقی شامل یک سلسله مراتب از سطوح سـاختاری هسـتند و بـرای زیسـتن، درک و تجدید این سطوح، تنـوع گسـترده‌ای از اِشـکال هـوش را کـه تـا انـدازه‌ای بـه‌خوبی درک شـده‌اند، پیش‌فرض می‌گیرند (خلج، ۱۳۸۳، ص۱۳۰). سطوح رفتار اخلاقی و اِشـکال هـوش مـرتبط مـی‌توانـد به صورت زیر تبیین شود:

- روابط و فعالیت‌های اخلاقی (رویه‌های اخلاقی یا اخلاق) در هـوش عملـی، هـوش عینـی، هوش تقلیدی؛
- جامعه اخلاقی ـ هوش بین‌فردی و ارتباطی؛
- وجدان اخلاقی ـ هوش هیجانی، هوش ارزیابی‌کننده؛
- علم اخلاق (اخلاق علمی مانند اخلاق زیستی، اخلاق فنـاوری و اخـلاق ماشـینی) ـ هـوش منطقی، هوش ریاضی و هوش فنی؛
- فلسفهٔ اخلاق (اخلاق و فرااخلاق) ـ هوش انتزاعی و هوش نظری؛
- معنویت اخلاقـی (درک، یـادگیری، ابـداع و اجـرای ارزش‌هـای اخلاقـی) هـوش توصـیفی، هوش بلورین، هوش روان و خلاق و هوش تفسیری.
- همان‌طور که واضح است، سیستم اخلاقـی بخـش جدایی‌ناپـذیر زنـدگی اجتمـاعی اسـت. برای زیستن در یـک سیسـتم اخلاقـی، مـا نیازمنـد اسـتفاده از ترکیبـات مختلـف از اشـکال

1. Bynum

گوناگون هوش هستیم. فیلسوفان و دانشمندانی که به دنبال راه‌های مفهومی و فنی برای اجرای یک کُد اخلاقی برای عوامل دارای هوش مصنوعی هستند، باید پیچیدگی زندگی اخلاقی را بپذیرند.

۳. ۱. وجدان اخلاقی

وجدان اخلاقی، یک وضعیت و فرایند اخلاقی پیچیده است. وجدان اخلاقی در سطح بالای ساختاریافته است، اما به صورت متوالی شکل گرفته و به طور نامتقارن رشد کرده است (Rey, 1924, p. 29). ساختار وجدان اخلاقی، افزون بر یک سری عادات، احساسات و حس‌ها، دیدگاه‌ها و باورها، شامل انعکاسات خاصی است که منشأ فلسفهٔ اخلاق و آغاز سطح معنوی زندگی اخلاقی را به عنوان سطوح سازنده و اجزا نمایندگی می‌کنند.

به عنوان سطح پایه رفتار اخلاقی، عادات اخلاقی از بیرون تشویق می‌شوند و توسط سیستم تنبیهات منفی و مثبت تحمیل می‌گردند. بنابراین، رفتار اخلاقی در طول زندگی در یک سیستم تعهد که احساسات منفی را برمی‌انگیزد، ادغام می‌شود و در نتیجه، تمام زندگی اخلاقی به عنوان زندگی در محدودیت ذهنی تجربه می‌شود.

گزینه‌ها و دیدگاه‌های اخلاقی اغلب نه از تجربیات شخصی، بلکه از طریق «شاگردی» از فرهنگ نسل‌های گذشته درونی می‌شوند و محتوای آنها، تکامل و نهایت آنها به ندرت درک می شود، بنابراین، بسیار پایدار، اما چندان انعطاف‌پذیر نیستند.

باورها، مجموعه‌های شناختی، عاطفی و ارزیابی هستند که می‌توانند در هر حوزهٔ فرهنگی در سطوح مختلف ساختاربندی شوند و نسبت این اجزا را در بر می‌گیرند. باورهای اخلاقی نسبتاً ثابت هستند؛ زیرا با منطق همراه و با عواطف تقویت می‌شوند. با این حال، باورهای ناسازگار می‌توانند همزمان وجود داشته باشند و اجرای متوالی یا جایگزین آنها ممکن است (علیزاده، ۱۳۹۷، ص۱۷۰). ثابت شده است که این اجزای وجدان اخلاقی می‌توانند به راحتی در زمینه‌ها و موقعیت‌های مختلف جابجا شوند (که می‌تواند یک مزیت باشد)، اما در عین حال، اغلب بسیار انعطاف‌ناپذیر و در مواقع دیگر صرفاً اظهاری و حتی نادرست هستند.

وجدان اخلاقی به عنوان یک جنبه از فعالیت انسانی، هم در سطح پایین و هم در سطح بالا (از عادات پایه تا زندگی معنوی)، در رفتار تجلّی می‌یابد و نه‌تنها شناخت و عاطفه را بلکه ارادهٔ آزاد را نیز بیان می‌کند. وجدان در اساس به معنای رفتار تعیین شده به طور خودکار، ظرفیت

اجتناب از فرمان و کنترل، توانایی برقراری ارتباط و همکاری و همچنین توانایی ایجاد اهداف، ابزارها و ارزش‌های جدید از طریق عمل است. بنابراین «وجدان»، یک منبع مهم خلاقیت را نمایندگی می‌کند و در عین حال، یک ویژگی هدف بسیار دشوار برای طراحان هوش مصنوعی به شمار می‌رود. با این حال، بسیاری از روان‌شناسان شناخته شده معتقدند که ناخودآگاه، منبع ماندگار و بی‌پایان خلاقیت است (The Unconscious, 1982, p. 4).

۳. ۲. شناخت اخلاقی

تحقیقات اخیر در حوزهٔ هوش مصنوعی که با مطالعات روان‌شناختی، روان‌تحلیلی و حتی روان‌آزردگی آغاز شد، به درک بهتر نمایش ساختارهای اجتماعی و موفقیت‌هایی در شناسایی، طبقه‌بندی و تجزیه و تحلیل وضعیت‌های هیجانی انسانی در شکل فعلی آن اجازه داده است (Arieti & Bemporad, 1978, pp. 78-80). تلاش شده است تا مناطق مغزی مسئول شناخت اجتماعی را مکان‌یابی کنند و ساختارهای شناختی که ویژگی‌های شناخت حوزه اجتماعی و دیگر روابط اجتماعی متنوع را آشکار می‌کنند، تعریف شوند.

با این حال، حتی نویسندگانی مانند سلومان[1] که مفاهیم احساسات را به عنوان عناصر سازماندهی و کنترل در عوامل توسعه می‌دهند، می‌پذیرند که احساس نه یک زیرسیستم خاص ذهن، بلکه یکی از ویژگی‌های نفوذکنندهٔ آن است. بنابراین، مشخص است که می‌توان احساسات مختلف با انواع متفاوت کنترل انگیزشی مرتبط هستند و استراتژی‌هایی را برای مقایسه و انتخاب انگیزه‌ها به عنوان روندهای تصمیم‌گیری برای سیستم‌های خودگردان که منابع محدود زمان و تلاش دارند، اما انگیزه‌ها و احساسات متعدد یا حتی متضاد را دارا هستند، توسعه داد (Sloman, 1990, pp. 237-238). روندهای تصمیم‌گیری بر اساس دیدگاه سلومان، بر معیارهای اخلاقی استوار است.

یک سلسله‌مراتب الگوهای ساختاری[2] و پویایی که توسط زبان اتخاذ شده است، به عنوان یک مُدل شناخت اجتماعی توسعه یافته است. این سلسله‌مراتب، دانش قوانین طبیعی یا اجتماعی یا استفاده از قواعد شناخته شده را پشت سر می‌گذارد و پیشرفت دانش را در شرایط

1. Sloman
2. Prototype

پیچیدگی، ابهام یا نقص اجازه می‌دهد (Churchland, 1998, p. 153).

متخصصان علوم رایانه افزون بر مسائلی مانند تشخیص و بازتولید اشکال، موضوعات مربوط به شناخت را به عنوان نمایش دیاگرامی غیرجمله‌ای دانش نیز بررسی کرده‌اند (عبدالهی، ۱۳۹۴، ص ۳۳). بنابراین، دانش فنی مبتنی بر روان‌شناسی شناختی و دیگر علوم شناختی می‌تواند مسائل روان‌شناسی فلسفی را حل کند.

شناخت اخلاقی در نگاه اول به عنوان یک جنبه از شناخت فلسفی باقی می‌ماند، اما دهه‌هاست که شناخت علمی را نیز در بر می‌گیرد و اکنون به شکل شناخت فنی ضمنی در اجرای یک کُد اخلاقی در حوزه‌ای مانند «اخلاق ماشینی» تبدیل شده است. شناخت اخلاقی علمی خود را به شکل شناخت تجربی، نظری و فرانظری نشان داده است. اشکال تجربی شناخت امروزه با چندرسانه‌ای پشتیبانی می‌شوند.

حتی اخلاق محاسباتی (اخلاق فلسفی) مورد بحث که تاریخ اخلاق و نظریه‌های اخلاقی اساسی را به طور جامع بررسی می‌کند، تا حدّی یک اخلاق تجربی است؛ زیرا بر افراد و مسائل اخلاقی آنها اعمال می‌شود، اسناد و متون قانونی را بررسی می‌کند، زندگینامه راهنمایان ارائه می‌دهد و یا مدل‌سازی موقعیت‌های اخلاقی مشکل را تشویق یا روندهای تصمیم‌گیری مناسب را پیشنهاد می‌کند.

بررسی و به کارگیری سطح روحانی سیستم اخلاقی پیچیده‌تر است. برای درک و ایجاد یک «اخلاق ماشینی»، ترکیب شناخت اخلاقی و فنی لازم است. شکل شناخت ترکیبی می‌تواند با بررسی تطبیقی اخلاق انسانی و ماشینی به طور جامع ایجاد شود. چنین مطالعه‌ای حتی می‌تواند دیدگاه تکامل همزمان آنها را به عنوان یک گزینه بدیل محتمل برجسته کند.

۳. ۳. معنویت اخلاقی

«معنویت» یکی دیگر از ویژگی‌های زندگی اخلاقی است، اما ضروری نیست. اگر چنین نبود، اخلاق یک حوزه از زندگی اجتماعی نمی‌شد. معنویت اخلاقی بر فرهنگ اخلاقی متکی است که اکثریت از آن محروم هستند. فرهنگ با تجربهٔ ذهنی برخی ارزش‌های خاص به یک سبک زندگی تبدیل می‌شود. بنابراین، اگر چه ارزش‌ها یک سطح ساختاری از سیستم اجتماعی را نمایندگی می‌کنند، اما نه یک حوزهٔ جداگانه، بلکه یک مؤلفه از هر جزء سیستم اجتماعی را تشکیل می‌دهند. (رمضانی و درخشی فیضی، ۱۳۹۴، ص ۳۷). بنابراین، هر واحد اجتماعی، محصول

عمل اجتماعی است. ارزش هم به شکل هدف و هم به صورت نتیجه بـرای افـراد عمـل مـی‌کنـد. عمل، حتی در ساختار درونی خود، ارزش را (بـه عنـوان معیـار انتخـاب بـین اشـیاء، نـوع انگیـزه، ابزارها یا شرایط، عنصر تأیید تصمیم، پایهٔ ارزیابی نتایج و هنجار برای فرصت شـروع مجـدد عمـل در مواقعی که ناسازگاری بین اهداف و نتایج وجـود دارد) دربـر مـی‌گیـرد. ارزش وجـود دارد؛ زیـرا عمل را آغاز می‌کند.

هر زیرسیستم جامعه توسط نوع مربوطه عمل اجتماعی که توسط سیستم ارزشـی خـاص خـود (اقتصادی، فنی، سیاسی یا اخلاقی) که حوزهٔ فرهنگـی خـاص را تشـکیل مـی‌دهـد، هـدایت مـی شود. فرهنگ اخلاقی یک سیستم ارزش‌های اخلاقی است که در اطـراف یـک ارزش مرکـزی کـه می‌تواند برای تمام زمان‌ها و مکان‌هـای اخلاقـی مشـترک باشـد، سـازماندهی شـده اسـت. دیگـر ارزش‌های اخلاقی می‌توانند محتـوا و اهمیـت خـود را از طریـق تغییـرات در سیسـتم‌های مرجـع تغییر دهند. تمام ارزش‌های اخلاقی از زوایای مختلف توسط نظریه‌های اخلاقی که بُعـد فرهنگـی معنویت اخلاقی را نمایندگی می‌کنند، تفسیر می‌شوند.

۳. ۴. ارزش‌های اخلاقی

اگر تلاش کنیم یـک کُـد اخلاقـی بـرای ماشـین‌ها (و احتمـالاً انسـان‌ها) ایجـاد و اعمـال کنـیم، می‌توانیم کار خود را همانند هر حوزه عمل دیگری بر ارزش‌های خاصی متکی سـازیم. بهـره‌وری، ارزش پایه‌ای برای هر نوع عمل فنی است و مختص ایجـاد اخـلاق ماشـینی (و بنـابراین، بهـره‌ور) نیست. آیا استناد به ارزش‌های اخلاقی در حـین سـاخت یـک اخـلاق جهـانی جدیـد، کارآمـد و مناسب برای دنیای مصنوعی (فنی، مجازی، فکری) ما، بـرای همـه مـا، شـهروندان، «شـهروندان شبکه»[1] یا «عقلای شبکه»[2] مفید خواهد بود؟ موارد زیر قابل مشاهده است:

- ارزش‌های اخلاقی عام، ترکیبی، مبهم یا نامشخص هستند و در حال تکامل هستند،
- کسب هر ارزش اخلاقی نیازمند استفاده و پرورش یک یا چند توانایی/ ویژگی روحانی است:
- هوش، امکان ارزیابی توانایی ما در انجام کار درست را فراهم می‌کند؛
- تخیل، از اقدامات بالقوه بد جلوگیری می‌کند؛

1. netizen
2. intellizen

- نتیجه، می‌تواند عادت صداقت را بیابد؛
- اراده، می‌تواند به فرد در انجام وظیفه‌اش کمک کند؛
- ظرفیت تلاش، حفظ آزادی اخلاقی را تسهیل می‌کند؛
- توجه، ظرفیت یادگیری و بصیرت می‌تواند احترام را تضمین کند؛
- عینیت یافتن ارزش‌های اخلاقی همچنین توسط ویژگی‌های شخصیتی و نگرش‌های فرهنگی نیز شرطی می‌شود؛
- کسب ارزش اخلاقی مرکزی، دستیابی به دیگر ارزش‌های اخلاقی را نیز ضروری می‌سازد؛
- ارتباط بین ارزش‌ها و هنجارها یک‌به‌یک نیست؛
- گسستگی میان ارزش‌ها و هنجارها، در ابهام هنجارها ادامه می‌یابد؛ اینها صرفاً الزامات دارای محتوایی هستند که توسط زندگی واقعی اخلاقی و زندگی عینی تعیین شده است؛
- هنجارهای اخلاقی همچنین مرتبط با ویژگی‌های هر جامعه اخلاقی/ فرهنگی نیز هستند.

۳. ۵. هنجارهای اخلاقی

حلِّ مشکلات مفهومی به توانایی ما در درک پدیدهٔ اخلاقی در یکپارچگی اجزای آن به معنویت، هنجارها و اثربخشی بستگی دارد (مختاری‌پور و سیادت، ۱۳۸۸، ص ۴). با این حال، شکاف بین شناخت و عمل، ارزش‌ها و هنجارها می‌تواند با تغییراتی که در درون خود نظریه (اخلاقی) رخ می‌دهد، تخفیف یابد. اگر نظریهٔ نظام‌بخشی دارای نتایج تجربی باشد؛ هنجار می‌تواند به عنوان یک گام عملی و فکری مورد انتظار در فرایند تبدیل گزاره‌های ادعایی به گزاره‌های قاعده‌مند و استانداردسازی بی‌پایان فعالیت‌های انسانی تعیین شود. در این صورت، هنجارها ادامه دستاوردها، شناخت‌های قاعده‌مند و حتی فرامینی هستند. نظریه‌ها با استفاده از هنجارهای کاربردی حتی مشتق شده از نظریه، کارکردی می‌شوند و اگر بتوان یک مُدل کارکردی را با جزئیات هنجارهایی که در اساس نظریه ساخته و آزمایش کرد، نظریه اثبات می‌شود.

فرهنگ اخلاقی، بخشی از فرهنگ عمل به همراه فرهنگ فنی، سیاسی یا حقوقی است و دارای خصلت هنجاری قوی است (عبدالهی، ۱۳۹۴، ص ۳۸). هنجارهای اخلاقی ویژگی‌های خاص شناخته شده‌ای دارند، اما در این زمینه بسیار مهم می‌شوند. 'عامل‌های اخلاقی ما، با اجتناب از به کارگیری توانایی‌هایی که این ویژگی‌ها مفروض می‌گیرند، در واقع، ممکن است تنها ربات‌های معمولی باشند که تنها برای یک گروه کاری بسیار محدود و خیلی دقیق تعریف شده مفید باشند.

هنجارهای اخلاقی برای هر نوع عامل انسانی موجود است: تمام فعالیت‌های (انسانی) یک بعد اخلاقی دارند و باید از این منظر ارزیابی/ قضاوت شوند. اخلاق یک ضرورت اجتناب‌ناپذیر برای تمام فعالیت‌های (انسانی) است، اما اعمال هنجارهای اخلاقی در حوزه‌های مختلف پیچیده است و تا کنون تنها نیازمند استفاده از توانایی‌های انسانی و کسب برخی ابزارهای فرهنگی است (کریمی واقف و عبدخدایی، ۱۴۰۰، ص ۸۹) هنجارهای اخلاقی عام، جایگزین‌پذیر و پایدار، نه‌تنها با پرورش و استفاده از ویژگی‌های فکری، بلکه معنوی و خلاق نیز می‌توانند به درستی تفسیر و به طور مؤثر استفاده شوند.

بنابراین، می‌توان از مهم‌ترین دشواری‌های ایجاد رفتار اخلاقی برای عامل‌های انسانی و مصنوعی بحث کرد. در حالی که اخلاق انسانی ناکارآمدی‌های درونی خود ارائه می‌دهد؛ اخلاق انسانی از نظر نظریه‌های اخلاقی قابل اصلاح نیست. به عنوان مثال، اخلاق اکثریت نمی‌تواند شکست‌های نظریه‌های فردگرایانه یا جمع‌گرایانه را از بین ببرد. یک ناسازگاری دوگانه عمیق بین سطوح نظری و عملی اخلاق انسانی و سطوح رفتاری و روانی اخلاق جداگانه تکامل می‌یابند. اخلاق تقریباً تنها به عنوان اخلاق وظیفه‌گرا حرفه‌ای باقی می‌ماند. بنابراین، اخلاق انسانی نمی‌تواند مُدلی برای اخلاق ماشینی باشد؛ زیرا برخی از اشکال هوش مورد نظر مطالعه نشده (مانند هوش ارتباطی و ارزیابی)، برخی کمتر مطالعه شده، مثل هوش تفسیری و یا برخی اشکال هوش مصنوعی رویکردهای جدیدتری دارند، مثل هوش عاطفی.

هوش اخلاقی به عنوان یک پیش‌نیاز برای اجرای یک کُد اخلاقی در نظر گرفته نمی‌شود و مورد مطالعه قرار نمی‌گیرد (خزاعی و زمان‌فشمی، ۱۳۹۲، ص ۱۱). توسعهٔ اَشکال خاص یا ویژه هوش به عنوان یک جایگزین کاربردی برای یک فرم انتزاعی از هوش که حتی یک فرم عمومی هوش نیست، اولویتی برای تحقیقات هوش مصنوعی نیست. اگر چه می‌توان مشاهده کرد که بر این موضوع اصرار می‌شود، این اَشکال ویژه هوش تنها می‌تواند سرعت، قابلیت مدیریت و جهان‌شمولی را به دست آورد، اما ممکن است این دستاوردها را به عنوان از دست دادن عمق یا ظرفیت تعیین سرنوشت خود و تمایز تفسیر کرد.

۴. کشف هوش اخلاقی، گامی به سوی اجرای قواعد اخلاقی

با بررسی ساختار پیچیده سیستم اخلاقی، می‌توانیم دشواری‌های اجرای یک کُد اخلاقی را روشن‌تر مشاهده کنیم. همچنین با تجزیه و تحلیل هوش (انسانی)، می‌توانیم احتمالات ایجاد

گونه‌های کارآمد هوش مصنوعی را به طور واقع‌بینانه‌تری ارزیابی کنیم.

عامل‌های مصنوعی کارآمد باید نه‌تنها از نظر فنی، بلکه بر اساس «ارزش‌ها و هنجارهای اخلاقی»[1] نیز با «توانایی‌ها»[2] مهارت‌ها و در صورت امکان نه‌تنها با عواطف یا چیزهای ساده (اما مصنوعاً دشوار برای بازآفرینی)، بلکه با «یک زندگی معنوی»[3] نیز مجهز شوند.

در حوزه تخصصی روان‌شناسی، ساختار «هوش» تجزیه و تحلیل شده است. توجه نابرابری به اشکال مختلف هوش نشان داده شده است، ابتدا آنها که در دسترس‌تر (قابل تجزیه و اندازه‌گیری) بودند، متمایز شدند؛ طبقه‌بندی‌ها بر اساس معیارهای در حال تغییر انجام شده است. بنابراین، برخی اَشکال هوش (مانند هوش ریاضی، زبانی یا بین فردی) در چندین رده ظاهر می‌شوند و شکل‌های مختلف؛ مانند انواع مختلف طبقه‌بندی شناسایی، تقلیدی، تفسیری و خلاق با یک‌دیگر مرتبط هستند، اما بسیاری از اشکال واقعی و مهم هوش به طور کامل نادیده گرفته می‌شوند. نمونه‌ها می‌تواند شامل آنهایی باشد که در سطوح بسیار انتزاعی و پیچیده از هر نوع فعالیت استفاده و پرورش می‌یابند و همه آنها به سادگی به عنوان «بیان‌های هوش خلاق»[4] ارزیابی می‌شوند.

به این نقایص نظری، شکست‌های آموزشی از نبود هدف خاصی برای ایجاد و پرورش هوش همراه می‌شوند. در نتیجه، اگر چه هر انسان سالم به طور بالقوه دارای تمام ظرفیت‌های هوشی است، اما افراد نمی‌توانند تمام اَشکال هوشی که به آنها ارائه می‌شود را بشناسند، ارزیابی یا استفاده کنند.

در اینجا سه طبقه‌بندی جدید، غیر روان‌شناختی اما فلسفی از اشکال هوشی را که همگی به بررسی و پرورش هوش اخلاقی به عنوان شرط وجود فرهنگی عامل‌های انسانی و مصنوعی نفوذ می‌کنند، قابل پیشنهاد است:

- معیار پیشنهادی اول، می‌تواند توسط حوزه‌های فرهنگی که توانایی‌های مختلف انسانی در آنها پرورش می‌یابد، آموزش داده می‌شود و تجلّی می‌کند، نمایندگی شود. بر اساس این معیار، ما می‌توانیم اَشکال علمی، هنری، فنی، سیاسی یا اخلاقی هوش را تمایز قائل شویم.

1. moral values and norms
2. faculties
3. spiritual life
4. expressions of creative intelligence

- معیار دوم، وضعیتی عملی دارد. هوش بر اساس نیازها، هنجارها و ارزش‌های مختلف حوزه‌های فعالیت شکل می‌گیرد. ارزش‌های اخلاقی یکی از سیستم‌های ارزشی را که در هر سازمان اجتماعی موجود است، تشکیل می‌دهند و تحقّق آنها ایجاد، توسعه و کاربرد یک مجموعه توانایی‌های (انسانی) مانند هوش اخلاقی را ضروری می‌سازد. همان‌طور که مشخص است، دیگر توانایی‌های مورد نیاز برای دستیابی به هر ارزش اخلاقی می‌توانند تعیین شوند و متقابلاً، توسعه هر کیفیت درونی یا اکتسابی؛ فعالیت‌ها، ابزارها و محیط‌های خاصی را ایجاب می‌کند.
- استراتژی نظری سوم برای یکپارچه‌سازی هوش اخلاقی در یک سیستم توضیحی منسجم، بیشتر جمع‌آوری تمام ویژگی‌های ضروری، خاص، ثابت تعیین‌کننده آن در میان تمام اشکال پیچیده دیگر هوش است تا متمایز کردن آن. بنابراین، ما فکر می‌کنیم هوش اخلاقی در سطوح و شدت‌های متغیّر، تمام اشکال دیگر هوش انسانی را در بر می‌گیرد. آن با اشکال شکل‌گیری و تجلی‌اش، به تمام حوزه‌ها و فرایندهای فکری نفوذ می‌کند.

هوش اخلاقی، یک ترکیب از تجلیات عینی دیگر اشکال هوش است. هوش اخلاقی یک شکل انتزاعی از هوش نیست، حتی یک شکل خاص از هوش هم قلمداد نمی‌شود (مومن‌نژاد، ۱۳۸۳، ص ۷۰) آن نه شکل عام و نه شکل خاص (انتزاعی یا عملی) هوش است. تا کنون این شکل از هوش مورد مطالعه قرار نگرفته است. بنابراین، متأسفانه همچنین ناکافی پردازش شده است.

هوش اخلاقی نباید با هوش عمومی (گرایش کارکردی فکری عمومی) اشتباه گرفته شود و درجه تجلّی این نوع هوش می‌تواند بسیار متفاوت با امکاناتی باشد که سطح توسعهٔ اولی ارائه می‌کند. هوش اخلاقی در اساس در زمینه‌های خاص (اخلاقی) فعّال و مؤثر است، اما این گونه موقعیت‌ها در تمام حوزه‌های فعالیت بروز می‌کنند و هر انسانی در هر سطحی تجربیات اخلاقی دارد. هوش اخلاقی مانند برخی اَشکال هوش هنری (مانند هوش بصری، حرکتی یا موسیقی) و انواع هوش علمی (ریاضی، زبانی، تحلیلی یا ترکیبی، نظری یا عملی) یک شکل ویژه از هوش نیست.

آمیختن هوش اخلاقی به یکی از اَشکال هوش مورد بحث در بالا یا دیگری غیر ممکن است، بنابراین، نمی‌تواند از آنها جدا شود. با این حال، به عنوان یک شکل پیچیده و ترکیبی از هوش، هوش اخلاقی نه‌تنها به آن اَشکال و سطوح هوش مورد بحث، بلکه به نظر می‌رسد اساساً به سطح تکامل آگاهی اخلاقی و وجود اجزای تجزیه شده، درجهٔ وابستگی متقابل و کارکردی آنها

نیز بستگی دارد.

با این حال، اجرای یک کُد اخلاقی در یک ماشین می‌تواند با استفاده از نتایج کارکرد برخی جنبه‌های مشترک هوش کلامی و عددی انسان و ظرفیت استدلال و حافظه تسهیل شود؛ زیرا هر عمل موفق (انسانی) در مجموع نتیجه فعالیت روح است. افزون بر این، تعیین عوامل کافی برای تعریف کارکردهایی که می‌توانند توانایی‌های رفتاری پیچیده و نگرش‌های فرهنگی مشخصه رفتار اخلاقی را شبیه‌سازی کنند، ضروری خواهد بود.

بنابراین، ما مجبور نیستیم یک کُد اخلاقی را اجرا کنیم، بلکه می‌توانیم برای ایجاد هوش اخلاقی، نه ایجاد یک واقعیت ثابت، بلکه شرایط یک پتانسیل را آرزو کنیم. برای انسان‌ها، هوش بالقوه از هوش بلوری مهم‌تر است. کسب یک شکل پیچیده از هوش لزوماً ماشین را به انسان‌ها نزدیک نمی‌کند، اما اخلاق ماشینی می‌تواند در غلبه بر برخی دشواری‌های اخلاق انسانی کمک کند؛ به طوری که مستقیماً از یک نظریهٔ اخلاقی استخراج شود، با تکنیک‌های فکری پشتیبانی شود و بر ارزیابی عینی احتمالات مربوط به الزامات متکی باشد. با ابزارهای فنی که دقت، شفافیت و کارایی را تضمین می‌کنند اجرا شود و با فناوری‌های مبتنی بر دانش و رباتیک شناختی به دست آید، اما مسائل مولّد، شرطی‌ساز و کنترل‌کننده باور، در نهایت، در سطح روحانی وجدان اخلاقی مصنوعی ظاهر خواهند شد.

درک بهتر اخلاق انسانی به این نتیجه می‌رساند که اخلاق ماشینی نه شبیه‌سازی اخلاق انسانی، بلکه نتیجه یک کشف اخلاقی مصنوعی خواهد بود. با این حال، این اخلاق حتی می‌تواند برای بهبود اخلاق انسانی به عنوان یک بیان دیگر از گرایش به مصنوعی‌سازی کنونی مناسب باشد.

۵. ویژگی‌های عوامل دارای هوش مصنوعی[1]

ماهیت اخلاق انسانی و عدم تناسب اساسی میان اخلاق انسانی و اخلاق ماشینی که از برخی ویژگی‌های عوامل دارای هوش مصنوعی ناشی می‌شود، سبب می‌شود که اخلاق انسانی نتواند مُدل کارآمدی برای اخلاق ماشینی باشد (London, 1997, p. 61). به دلیل سطوح پیچیدگی، کارکردهای خاص و در اساس نیاز به درجه‌ای از آزادی، عوامل اخلاقی مصنوعی باید به عنوان

1. AI Agents

موجودات فـردی مجهـز بـه دیگـر ویژگی‌هـای ضـروری در نظـر گرفتـه شـوند. عوامـل اخلاقـی مصنوعی بایستی دارای ویژگی‌های زیر باشد:

- پیچیده، تخصصی، خودگردان یا خودکفا، با رفتارهای پایدار و حتی غیرقابل پیش‌بینی باشند؛
- سیسـتم‌هایی بـا عملکـرد رفتـاری بـاز و حتـی آزادانـه، دارای مکانیزم‌هـا و رونـدهای تصمیم‌گیری خاص، انعطاف‌پذیر و شهودی باشند؛
- موجودات فرهنگی باشند؛ زیرا رفتار آزادانه، ارزش فرهنگی را به اعمال یک موجـود طبیعـی یا مصنوعی می‌افزاید؛
- سیستم‌هایی باز نه‌تنها برای آموزش، بلکه برای تربیت نیز باشند؛
- موجوداتی با «دانش زندگی» باشند؛
- موجوداتی با انواع مختلف و حتی چندگانه هوش مانند هوش اخلاقی باشند؛
- نه‌تنها با اتوماسیون‌ها، بلکه با باورها (مجموعه‌های‌شناختی و عاطفی) مجهز شده باشند؛
- بتوانند بیندیشند و بازتاب کنند؛ زیرا زنـدگی اخلاقـی، نـه‌تنها یـک فعالیـت آگـاهی، بلکـه همچنین یک شکل از معنویت است؛
- به عنوان عناصر/اعضای برخی جوامع واقعی (فیزیکی یا مجازی) رفتار کنند.

اخلاق ماشینی مجبور است مسئله دشواری را حل کند که نیاز نظـری را بـا الـزام عملـی در ارتبـاط بـا آزادی اخلاقـی ترکیـب مـی‌کنـد. اخـلاق در رفتـار انسـانی، اجـرای عملـی آزادی اخلاقـی را پیش‌فرض می‌گیرد (خزاعی و زمان‌فشمی، ۱۳۹۲، ص ۷). رفتار انسـانی تنهـا بـا ایـن پیش‌فـرض و اعمـال هنجارهای اخلاقی می‌تواند ارزش فرهنگی کسب کند. افزون بر این، هنگامی که ماشـین مـد نظـر است، رابطه آزادی ـ مسئولیت باید هم در یک معنای گسترده و هم در یـک معنـای محـدود عمـل کند؛ زیرا مسئولیت، شرط آزادی است و آزادی، علت مسئولیت است. به طـور صـریح تـر، درجـهٔ مسئولیت نه‌تنها به انواع و سطوح آزادی بستگی دارد، بلکه توسط آنها تعیین می‌شود.

بنابراین، اخلاق ماشینی باید نیاز به فـراهم کـردن آزادی انتخـاب در حوزه‌هـای عمـل و شـرایط تعیین شده را مورد توجه قرار دهد. انسان بایـد ریسـک‌های ناشـی از اعطـای آزادی بـه ماشـین‌ها را بپذیرد و احتمال افزایش درجهٔ آزادی ماشین‌ها را نیز قبـول کنـد. رفتـار وابسـته نـه‌تنها بـه انسـان‌ها، بلکه به تصمیم خودش و حتی تصمیمات دیگر ماشین‌ها نیز هست.

اعمال ویژگی‌های ذکر شده لزوماً تلاش برای طراحی، ساخت و آموزش یک ماشـین بـه عنـوان یک موجود نیمه‌انسانی را ایجاب نمی‌کند. از طرف دیگر، حتی رفتـار انسـانی نیـز تنهـا بـا ایجـاد و

اجرای یک اخلاق که از نظر فلسفی برقرار شده و از نظر علمی استنتاج شـده اسـت، مـی‌توانـد بـه کمال برسد (Ey, 1998, p. 73). این یک اخلاق مصنوعی مشترک و بهتـر خواهـد بـود کـه مـی‌توانـد توسط انسان‌ها و ماشین‌هایی که در میانه طبیعی و مصنوعی قرار دارند، اجرا شود.

۶. هوش به عنوان یک سیستم با ساختار داخلی

برای دستیابی به هر ارزش اخلاقی، رفتار اخلاقی عوامل دارای هـوش مصـنوعی و همچنـین رفتـار اخلاقی انسان، نه‌تنها به هوش، بلکه به بسیاری از ویژگی‌های روانی متنوع نیـز نیـاز دارد. بـه هـوش خود، در شکل‌های خاص آن در زمینـه‌های مختلـف فعالیـت انسـانی یـا مصـنوعی نیـاز اسـت و هوش می‌تواند از طریق تحلیل درونی پرورش و توسعه یابد.

هوش انسانی بر اساس فعالیت انسانی متنـوع شـده اسـت. فرضـیه مـا در مـورد وجـود هـوش اخلاقی و کارکرد آن، با به کارگیری روش‌شناسی سیستمیک در روان‌شناسی مجـاز شـده و بـا یـک چشم‌انداز فلسفی یکپارچه‌کننده شکل‌های فرهنگی ادامه می‌یابد.

تحقیقات در مورد ساختار هوش انسانی در دو جهت پیش رفته است که هر یـک نقشـه ناقصـی از نمایش خود را به جـا گذاشـته اسـت. بـه صـورت اسـتقرایی، اشـکال هـوش ریاضـی، زبـانی، توصیفی، تفسیری و نظری بررسی شده‌اند، اما در مورد هـوش علمـی کـار نشـده اسـت. فهرسـت هوش ادبی، موسیقی و پلاستیک تهیه شده است، اما فهرست هوش هنـری تهیـه نشـده اسـت. بـه روش قیاسی، یک کارآیی کارکردی فکری عمومی (هوش عمومی) تعیین شده است.

هوش فنی به عنوان یک شکل از هوش عملی تجزیـه و تحلیـل شـده اسـت. هـوش اخلاقـی و سیاسی نیز عمدتاً عمل‌گرا و به شـدّت توسـط هنجارهـا کنتـرل مـی‌شـود، امـا توسـط ارزش‌هـای خاصی که دنبال می‌شوند و ابزارهایی که استفاده می‌شوند، متمایز می‌گـردد. بـا ایـن حـال، هـوش اخلاقی به راحتی در گروه اعمال عملی ادغام نمی‌شـود (مـومن‌نـژاد، ۱۳۸۳، ص ۵۱) بنـابراین، مـا دارای یک هوش عمومی که سطح فکری خاص هر رفتار انسانی را فراهم می‌کند، سـپس اشـکال خـاص هوش که انواع انتزاعی و عملی را در بـر می‌گیـرد و در نهایـت، اشـکال خـاص هـوش متمرکـز بـر ارزش‌های خاص، فناوری‌های آموزشی و تجربیات در محیط‌های مناسـب کـه توسـط زمینـه‌های عمل مختلف تولید می‌شوند، هستیم.

مرحلهٔ واقعی و سرعت پیشرفت در تحقیقات هوش مصنوعی، اساسـاً توسـط گـرایش آنهـا بـه

توسعه یک شکل انتزاعی از هوش تعیین می‌شود. برای اجرای کارآمد یا موفق یک قاعدهٔ اخلاقی، به یک شکل پیچیده از هوش انسانی و مصنوعی نیاز است.

۷. دشواری‌های اعمال قواعد اخلاقی برای عوامل انسانی و واجد هوش مصنوعی

آنها که با مشکلات دشوار و پیچیده در زمینه‌های مفهوم‌سازی و اختراع روبه رو هستند، تنها دانشمندان رایانه و محقّقان هوش مصنوعی نیستند. فیلسوفان و نمایندگان علوم انسانی نیز بر اساس چشم‌اندازهای محیط دانشی موجود، به باز بنیانگذاری ضروری حوزه‌های تحقیقاتی خود و بازسازی ساختاری آنها علاقه‌مند هستند. بازبینی جدّی بسیاری از مدل‌های تفسیری و تأویلی معتبر علوم اجتماعی و تأمل، متناسب با نیازهای فعلی دانش علمی و فنی، ضروری است. در عمل، اخلاق انسانی یک اخلاق انتخاب محدودیت و بی‌مسئولیتی است؛ اخلاق انسانی به عنوان یک نظریهٔ اخلاقی، مجموعه‌ای از تناقضات درونی ارائه می‌دهد (Bedau, 1998, pp. 55-59). بنابراین، می‌توان گفت اخلاق انسانی می‌تواند به عنوان یک مُدل برای اخلاق ماشینی عمل کند. اکنون زمان آن رسیده است که یک پایه علمی و فنی مشترک برای یک اخلاق مصنوعی کاملاً اختراع شده که بتواند توسط هم انسان‌ها و هم ماشین‌ها اعمال شود. متخصصان علم رایانه و دانشگاهیانی که خود را وقف تحقیقات هوش مصنوعی کرده‌اند، تلاش می‌کنند یک «کد اخلاقی» را به شرح زیر اجرا کنند: (Bedau, 1998, p. 73-77).

- بدون توجه به پیچیدگی واقعی فرهنگ اخلاقی، سطح معنوی سیستم اخلاقی و محیط واقعی فعالیت‌های سیستم هوش در نظر گرفته شود؛
- الزام توسعه یک هوش مصنوعی انتزاعی؛
- پیش‌فرض قرار دادن بهره‌وری و اختصاصی بودن، کفایت عوامل، انگیزه‌ها، اهداف، شرایط، ابزارها، راهبردها، ارزیابی‌ها، نتایج و ارتباطات پویای آنها در یک سیستم سایبری؛
- اجتناب از اختصاصی بودن شکل هوشی که می‌تواند رفتار مصنوعی کارآمد را در یک زمینه اخلاقی فراهم کند.

بنابراین، لازم است یک مطالعه در مورد هوش اخلاقی به عنوان پاسخی به مسئلهٔ پیچیدگی و اختصاصی بودن رفتار و فرهنگ اخلاقی با تحلیل جامع سیستم اخلاقی بر اساس ویژگی‌های واقعی آن انجام شود.

۸. موانع پیشِ روی هوش مصنوعی اخلاقی: تحلیل چالش‌های فنی، فلسفی و حقوقی

هوش مصنوعی (AI) ظرفیت عظیمی برای ارتقای زندگی بشر در زمینه‌های مختلف از جمله مراقبت‌های بهداشتی، حمل و نقل و آموزش دارد. با این حال، این پیشرفت‌ها با چالش‌های اخلاقی متعددی نیز همراه است. هوش مصنوعی اخلاقی به عنوان مجموعه‌ای از اصول و رویه‌ها برای اطمینان از این‌که هوش مصنوعی به طور مسئولانه و اخلاقی توسعه و استفاده شود، مطرح شده است.

الف) چالش‌های فنی

- **تعریف و اندازه‌گیری اخلاق:** چالش اصلی در هوش مصنوعی اخلاقی، تعریف و اندازه‌گیری مفاهیم اخلاقی مانند عدالت، انصاف و عدم تبعیض است. الگوریتم‌های هوش مصنوعی باید به‌گونه‌ای طراحی شوند که این ارزش‌ها را در تصمیم‌گیری‌های خود لحاظ کنند.
- **تضمین شفافیت و قابلیت تفسیر:** الگوریتم‌های هوش مصنوعی اغلب پیچیده و غیرقابل درک هستند که می‌تواند به عدم شفافیت در تصمیم‌گیری‌های مبتنی بر هوش مصنوعی منجر شود. لازم است الگوریتم‌ها به‌گونه‌ای طراحی شوند که قابل درک و تفسیر باشند تا بتوان در مورد اخلاقی بودن آنها قضاوت کرد.
- **مدیریت داده‌های سوگیرانه:** داده‌هایی که برای آموزش الگوریتم‌های هوش مصنوعی استفاده می‌شوند، می‌توانند مغرضانه باشند و این سوگیری‌ها را به تصمیم‌گیری‌های الگوریتم‌ها منتقل کنند. لازم است از روش‌های مختلفی برای شناسایی و حذف سوگیری‌ها از داده‌ها استفاده شود.

ب) چالش‌های فلسفی

- **ماهیت آگاهی و احساس:** هوش مصنوعی به طور فزاینده‌ای پیچیده می‌شود و این سؤال را مطرح می‌کند که آیا ماشین‌ها می‌توانند آگاه و دارای احساس باشند یا خیر. اگر ماشین‌ها آگاه باشند، از حقوق اخلاقی مشابه انسان‌ها برخوردار خواهند بود.
- **مسئولیت اخلاقی:** در صورت بروز مشکل با هوش مصنوعی، چه کسی مسئول خواهد

بود؟ آیا سازندگان، توسعه‌دهندگان، اپراتورها یا کاربران هوش مصنوعی مسئول خواهند بود؟

- **آزادی و استقلال انسان:** هوش مصنوعی می‌تواند به طور قابل توجهی بر زندگی انسان‌ها تأثیر بگذارد و این سؤال را مطرح می‌کند که آیا این امر به از دست رفتن آزادی و استقلال انسان می‌انجامد؟

ج) چالش‌های حقوقی

- **نیاز به قوانین و مقررات جدید:** در حال حاضر، هیچ چارچوب حقوقی جامعی برای هوش مصنوعی وجود ندارد. لازم است قوانین و مقرراتی برای تنظیم توسعه و استفاده از هوش مصنوعی به منظور محافظت از حقوق افراد و جامعه تدوین شود.
- **مسائل مربوط به حریم خصوصی و امنیت داده‌ها:** هوش مصنوعی نیازمند حجم عظیمی از داده‌های شخصی است که می‌تواند به مسائل مربوط به حریم خصوصی و امنیت داده‌ها منجر شود. لازم است ضمانت‌های لازم برای محافظت از داده‌های افراد در برابر سوء استفاده فراهم شود.
- **مالکیت معنوی:** اختراعات و نوآوری‌هایی که در هوش مصنوعی ایجاد می‌شوند، می‌توانند مسائل مربوط به مالکیت معنوی را مطرح کنند. لازم است قوانین مالکیت معنوی برای انعکاس پیشرفت‌های هوش مصنوعی به روز شوند.

راهکارها:

- **توسعه چارچوب‌های اخلاقی:** لازم است چارچوب‌های اخلاقی برای هوش مصنوعی تدوین شود که اصول و رویه‌های لازم برای توسعه و استفاده مسئولانه از هوش مصنوعی را مشخص کند.
- **افزایش شفافیت و قابلیت تفسیر:** الگوریتم‌های هوش مصنوعی باید به‌گونه‌ای طراحی شوند که قابل درک و تفسیر باشند تا بتوان در مورد اخلاقی بودن آنها قضاوت کرد.
- **مدیریت داده‌های سوگیرانه:** لازم است از روش‌های مختلفی برای شناسایی و حذف سوگیری‌ها از داده‌ها استفاده شود.
- **تحقیق و توسعه در زمینه هوش مصنوعی اخلاقی:** لازم است تحقیقات و توسعه بیشتری در

زمینه هوش مصنوعی اخلاقی انجام شود تا به چالش‌های موجود در این زمینه رسیدگی شود.

- **ایجاد گفت‌وگوی عمومی:** لازم است گفت‌وگوی عمومی در مورد هوش مصنوعی اخلاقی ایجاد شود تا آگاهی عمومی در مورد این موضوع افزایش یابد و ذینفعان مختلف بتوانند در مورد چالش‌ها و راهکارها بحث کنند.

نتیجه‌گیری

پیاده‌سازی یک نظام اخلاقی در سیستم‌های هوش مصنوعی، نیازمند شکل خاصی از هوش انسانی و مصنوعی است که «هوش اخلاقی» نامیده می‌شود. عاملان هوش مصنوعی دارای ویژگی‌های منحصر به فردی هستند که نحوهٔ تعامل با آنها را تعیین می‌کند. اخلاق انسانی دارای نقایص ذاتی است و نمی‌تواند به طور کامل مُدل مناسبی برای اخلاق ماشینی باشد. با این حال، برای طراحی و پیاده‌سازی اخلاق در ماشین‌ها، نیاز به ابزارهای کامل نظری و عملی است، از جمله منطق ویژه، روان‌شناسی کارآمد و فنون تصمیم‌گیری ترکیبی.

اخلاق ماشینی می‌تواند از کیفیت برتری نسبت به اخلاق انسانی برخوردار باشد؛ زیرا از علوم، مُدل‌سازی و فناوری‌ها بهره می‌برد. با این حال، پیاده‌سازی هوش اخلاقی در هوش مصنوعی، مسائل نظری و عملی ساختاری، کارکردی و رفتاری را به همراه خواهد داشت که باید پیش‌بینی و برطرف گردند.

در نهایت، آینده اخلاق انسانی و مصنوعی باید به دقت مورد بررسی و برنامه‌ریزی قرار گیرد. توسعه و تکامل هوش اخلاقی در سیستم‌های هوش مصنوعی، آینده فناوری و جوامع انسانی را شکل خواهد داد. یکپارچه‌سازی اصول اخلاقی در این سیستم‌ها، تأثیر عمیقی بر نحوهٔ تعامل آنها با انسان‌ها و تصمیم‌گیری‌هایشان خواهد داشت. از این‌رو، مهم است که چارچوب‌های اخلاقی مناسب و سازگار با ارزش‌های انسانی طراحی و اعمال شوند.

افزون بر این، با پیشرفت فناوری هوش مصنوعی و گسترش حضور آن در زندگی روزمرّه، باید قوانین و مقررات حاکم بر این حوزه نیز تدوین شود. نظارت بر فرایند طراحی و آموزش سیستم‌های هوشمند اخلاقی و تعیین چارچوب‌های مسئولیت‌پذیری برای آنها، ضروری خواهد بود. از سوی دیگر، باید توجه داشت که هوش اخلاقی در هوش مصنوعی، تنها محدود به رعایت قوانین و اصول اخلاقی نیست، بلکه شامل ظرفیت‌های عاطفی، همدلی و توانایی درک زمینه‌های

فرهنگی و اجتماعی نیز هست. بنابراین، یکپارچه‌سازی این جنبه‌های انسانی در سیستم‌ها، گامی مهم در جهت تحقّق هوش مصنوعی واقعاً اخلاقی خواهد بود. در نهایت، آیندهٔ روشن اخلاق در هوش مصنوعی، نیازمند تلاش مشترک متخصصان فناوری، فلاسفه اخلاق، روان‌شناسان و دیگر ذینفعان جامعه است تا اطمینان حاصل شود که این فناوری پیشرفته، در خدمت منافع و ارزش‌های انسانی به‌کار گرفته می‌شود.

فهرست منابع

بلبلی قادیکلایی، سمیه؛ پارسا نیا، حمید. (۱۴۰۲). مروری نظام‌مند بر دلالت‌های اخلاقی استفاده از هوش مصنوعی در فناوری دیجیتال و نسبت ان با اخلاق شکوفایی، راهبرد اجتماعی و فرهنگی. ۲ (۴۸)، ۷۷۱-۷۹۸.

خزاعی، زهرا؛ زمان‌فشمی، ندا. (۱۳۹۲). روش ها و موانع پیاده سازی اخلاق کانتی در ماشین های هوشمند. پژوهش‌های اخلاقی. ۱(۱۳)، ۵-۳۲.

رمضانی، مجید؛ درخشی فیضی، محمد رضا. (۱۳۹۲). اخلاق ماشینی، چالش‌ها و رویکردهای مسایل اخلاقی در هوش مصنوعی و ابرهوش، اخلاق در علوم و فناوری. (۸)۴، ۳۵- ۴۳. 20.1001.1. 22517634.1392.8.4.4.7

عبدالهی، طاهره؛ ربیعی، علی؛ امینی، محمدتقی، (۱۳۹۴)، رابطه هوش اخلاقی با سرمایه اجتماعی، اخلاق در علوم و فناوری. ۳ (۱۰)، ۲۹-۴۰.

علیزاده، زهرا. (۱۳۹۷)، ارتباط هوش اخلاقی و هوش هیجانی با رفتار اخلاقی، اخلاق در علوم و فناوری. ۲۳ (۱۳)، ۱۶۷-۱۷۴.

کریمی واقف، نرگس؛ عبدخدایی، زهره. (۱۴۰۰). چالش‌های فرا روی کاربست اخلاق در ماشین‌های هوشمند با تمرکز بر رویکرد اصل‌گرایی اخلاق، پژوهش‌نامه اخلاق. ۳ (۵۱)، ۶۹-۹۲.

محرابی، نازیلا؛ خراشادی زاده، سحر؛ کریمیان، راحله. (۱۴۰۲). شناسایی مؤلفه‌های هوش مصنوعی در پیاده‌سازی مدیریت دانش، علوم و فنون مدیریت اطلاعات. ۳ (۹). ۳۵۱-۳۹۰.

محمدعلی خلح، محمد حسین. (۱۳۹۳). دریفوس و تاریخ فلسفی هوش مصنوعی، غرب‌شناسی بنیادی، ۱ (۳). ۱۰۳-۱۲۸.

مختاری‌پور، مرضیه؛ سیادت، علی. (۱۳۸۸). مدیریت با هوش مصنوعی، مجله تدبیر. ۲۰۵، ۱-۱۲.

مومن‌نژاد، آیدا. (۱۳۸۳). آگاهی، هوشمندی و هوش مصنوعی، مجله اطلاع‌رسانی، ۴ (۱۲)، ۴۹-۷۲.

Arieti S. & Bemporad J. R. (1978). *Severe and mild depression: the*

psychotherapeutic approach. New York: Basic Books.

Bedau, M. A. (1998). Philosophical Content and Method of Artificial Life. In Bynum, T. W. and Moor, J. H. (Eds.) *The Digital Phoenix: How Computers are Changing Philosophy*. Blackwell Publishers, Oxford.

Bynum, T. W. (1998). Global Information Ethics and the Information Revolution. In Bynum, T. W. and Moor, J. H. (eds.), *The Digital Phoenix: How Computers are Changing Philosophy*. Blackwell Publishers, Oxford.

Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral intelligence. *International Journal of Interactive Multimedia & Artificial Intelligence*. 6(3).

Danielson, P. (1998). How Computers Extend Artificial Morality. In Bynum, T. W. and Moor, J. H. (Eds.). *The Digital Phoenix: How Computers are Changing Philosophy*. Blackwell Publishers, Oxford.

Ey, H. (1982). Conştiin a (The Consciousness), Editura Ştiin ifică, Bucureşti. Gregory, R. (2000) *Viitorul creatorilor de inteligen ă (The Future of Mind-Makers)*. Editura Ştiin ifică, Bucureşti.

Floridi, L., & Cowls, J. (2023). Ethical Artificial Intelligence: Towards a Discipline Freed and Focused. *Nature Machine Intelligence*, 5(3), 251-256

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence*. Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hillis, W. D., (2001). Maşina care gândeşte (The Pattern on the Stone. The Simple Ideas that Make Computers Work), Editura Humanitas, Bucureşti,

Laruelle, Fr. (1990) *Th orie des identit es, fractalit generalis e ET philosophie artificielle*, P. U. F., Paris.

Narvaez, D., & Endicott, L. (2009). Nurturing character in the classroom. EthEx Series, Book 4: *Ethical skills*. University of Notre Dame.

Narvaez, D., & Lapsley, D. K. (2005). The psychological foundations of everyday morality and moral expertise. *Character psychology and character education*, 140-165.

Rey, A., (1924). Invention artistique, scientifique, pratique in Dumas, G. (ed.) *Traite de Psychologie*, Tome II (Les foundements de la vie mentale), Premier livre, Chapitre VI, Librairie Felix Alcan, Paris.

Russell, S. J., & Norvig, P. (2020). *Artificial intelligence: a modern approach* (4th ed.). Pearson.

Sloman, A. (1990). Motives, Mechanisms, Emotions. in M. Boden (ed.), *the Philosophy of Artificial Intelligence*, Oxford University Press.

Wallach, W. & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.