فصلنامه علمی-پژوهشی اخلاق‌پژوهی
سال هشتم • شماره اول • بهار ۱۴۰۴
Quarterly Journal of Moral Studies
Vol. 8, No. 1, Spring 2025

صورت‌بندی الگوی مطلوب در نسبت اخلاق با هوش مصنوعی با تأکید بر مبانی اخلاقی علامه مصباح یزدی (ره)

رضوان هامانی* | مصطفی مختاری**

doi 10.22034/ethics.2025.51930.1777

چکیده

با وجود تحولات اساسی در زمینه‌های مختلف زندگی انسان که محصول پیشرفت‌های اخیر هوش مصنوعی است، آسیب‌پذیری جامعه و افراد در برابر تهدیدات این فناوری، به‌ویژه خودمختاری و استقلال روز افزون آن در تصمیم‌سازی‌ها، دغدغه‌های بسیاری در مورد پیامدهای اخلاقی در دو سطح طراحی و تعامل با آن ایجاد کرده و موضع‌گیری‌های متفاوتی در مواجهه با آن شکل گرفته است. در این تحقیق با روش توصیفی- تحلیلی با تأکید بر سطح «اخلاق با طراحی» به عنوان یکی از سطوح نسبت اخلاق با هوش مصنوعی که ناظر به مرحلهٔ طراحی و پیاده‌سازی نظریه‌های اخلاقی در ماشین است، می‌توان دیدگاه اخلاقی علامه مصباح را با تأکید بر مبانی و تمایزات خاصی که دارد ذیل نظریه «تعادل تأملی» و در سطح «اخلاق با طراحی» ذیل رویکرد «تلفیقی» در میان رویکردهای رایج موجود صورت‌بندی کرد. مبنای نظام اخلاقی ایشان بر پایه «غایت‌گرایی قُرب الهی» مستلزم این است که از اخلاق با طراحی برای ارتقای اخلاق انسانی و کسب کمال نهایی به عنوان یک ابزار مصنوع، استفاده کرد و اخلاق هوش مصنوعی را بر مبنای «غایت‌گرایی زمینه‌ساز کمال» صورت‌بندی کرد.

کلیدواژه‌ها

اخلاق هوش مصنوعی، غایت‌گرایی قُرب الهی، غایت‌گرایی، محمدتقی مصباح یزدی.

* دانشجوی دکتری، گروه فلسفه فیزیک، دانشگاه باقرالعلوم (ع)، قم، ایران. (نویسنده مسئول). | hamani@bou.ac.ir
** دانشجوی دکتری، گروه فلسفه اسلامی، دانشگاه باقرالعلوم (ع)، قم، ایران. | mokhtare.m14@gmail.com
تاریخ دریافت: ۱۴۰۴/۰۱/۲۵ ▣ تاریخ تأیید: ۱۴۰۴/۰۳/۱۰

■ هامانی، رضوان؛ مختاری، مصطفی. (۱۴۰۴). صورت‌بندی الگوی مطلوب در نسبت اخلاق با هوش مصنوعی با تأکید بر مبانی اخلاقی علامه مصباح یزدی (ره). فصلنامه اخلاق‌پژوهی. ۸(۲۶)، ۸۹-۱۱۲
doi: 10.22034/ethics.2025.51930.1777

۱. مقدمه

پیشرفت‌های کنونی در تحقیق، توسعه و کاربرد سیستم‌های «هوش مصنوعی»[1] نگرانی‌های گسترده‌ای را در مورد پیامدهای اخلاقی استفاده از هوش مصنوعی به وجود آورده است که در نتیجهٔ آن تعدادی از دستورالعمل‌های اخلاقی در سال‌های اخیر منتشر شده است. این دستورالعمل‌ها شامل اصول و توصیه‌های هنجاری فناوری‌های جدید هوشمند با هدف مهار پتانسیل‌های اخلال‌گر در اخلاق فردی و اجتماعی است (Hagendorf, 2020 , p. 99). هراس از تسلط ماشین‌های خودمختار بر آینده بشر، منجر به ظهور شاخه‌ای با عنوان اخلاق هوش مصنوعی شده است که به جنبه‌های اخلاقی در طراحی و تعاملات انسان و ماشین می‌پردازد. کاربرد فزاینده این فناوری ضرورت پرداختن به این دانش بین‌رشته‌ایی را بیش از پیش اهمیت بخشیده است. به طوری که روزالیند پیکارد، بنیانگذار و مدیر گروه تحقیقات MIT، بر این موضوع تأکید دارد که: «هر چه آزادی ماشین بیشتر باشد، به معیارهای اخلاقی بیشتری نیاز خواهد داشت» (Picard, 1995, p. 134).

واگذاری کنترل تصمیمات مهم به هوش مصنوعی مستلزم این است که افزون بر شرکت‌های پشتیبانی، توسعه‌دهندگان، طراحان و کاربران، برنامه‌ها و سیستم‌های هوش مصنوعی نیز در مراحل مختلف تحقیق، طراحی، پیاده‌سازی، توسعه و کاربرد به اصول و ارزش‌های اخلاقی مناسب مقید شوند. پژوهش در این زمینه ترکیبی از علوم کامپیوتر، علوم انسانی و فلسفهٔ اخلاق است. از این‌رو، دامنه تحقیقات نظری در این شاخه از چگونگی استفاده اخلاقی کاربران و تأکید بر طراحی و اجرای اخلاقی تا ساخت ماشین‌هایی که بتواند به شیوه‌ای خودمختار اخلاقی عمل کند را در بر می‌گیرد (Tolmeijer, 2020, p. 2). اهمیت پرداختن به الزامات اخلاقی تعامل انسان و ماشین غیرقابل‌انکار است، اما درک بهتر این ضرورت نیازمند یک رویکرد طبقه‌بندی شده است که در آن جایگاه شرکت‌ها و توسعه‌دهندگان، طراحان، کاربران و سیستم‌های خودمختار هوشمند در نسبت با الزامات اخلاقی روشن شود.

مسئله اخلاق در سیستم‌های هوشمند در سه شاخه اصلی که در فلسفهٔ اخلاق مطرح می‌شود، یعنی فرااخلاق، اخلاق هنجاری و اخلاق توصیفی، در اخلاق هوش مصنوعی نیز قابل بحث است. در شاخه فرااخلاق ضمن بررسی مفاهیمی همچون عاملیت اخلاقی ماشین‌های هوشمند، به این پرسش اصلی پرداخته می‌شود که آیا اصولاً یک ماشین هوشمند می‌تواند به عنوان یک عامل

1. Artificial intelligence (AI)

اخلاقی قلمداد شود؟[1] در بخش «اخلاق توصیفی» بدون هیچ قضاوتی تنها به توصیف جوامع و کشورها در مواجهه با اخلاق هوش مصنوعی پرداخته می‌شود. شاخهٔ «اخلاق هنجاری» به طرح امکان پیاده‌سازی اخلاق در ماشین اختصاص دارد. پرسش محوری این بخش آن است که چگونه می‌توان اخلاق را در ماشین پیاده‌سازی کرد؟ (عبدخدایی و دیگران، ۱۴۰۰، ص ۷۲). این بحث در میان اخلاق پژوهان هوش مصنوعی با عنوان «اخلاق با طراحی»[2] مطرح شده است. در نهایت «اخلاق کاربردی» نیز که در امتداد اخلاق هنجاری است، تعیین وظایف و بایدها و نبایدها و به عبارتی، هنجارهای اخلاقی هوش مصنوعی و چالش‌های اخلاقی که تا به امروز در مورد هوش مصنوعی مطرح بوده، مورد بررسی قرار می‌گیرد. یکی از پرسش‌های محوری این شاخه آن است که ماشین‌های هوشمند، به عنوان مصنوعات تکنیکی، چه چالش‌های اخلاقی‌ای را پیش رو قرار می‌دهند؟ مسئلهٔ سلاح‌های هوشمند، حریم خصوصی از جمله این چالش‌ها هستند.

در این بحث به امکان پیاده‌سازی اخلاق و نظریه‌های اخلاقی در ماشین و اینکه چه رویکردی در پیاده‌سازی اخلاق در ماشین‌ها مناسب و قابل اجراست پرداخته می‌شود. آنچه در این مقاله با عنوان «صورت‌بندی نسبت اخلاق با هوش مصنوعی» مورد توجه قرار گرفته است، ذیل شاخهٔ اخلاق هنجاری می‌گنجد. از این‌رو، لازم است به تفصیل بیشتری از زیرشاخه‌ها و دیدگاه‌های رایج در این شاخه و نسبت آن با هوش مصنوعی پرداخته شود. در این پژوهش با تبیین نظریه‌های اخلاقی مطرح در ماشین‌ها به دنبال پاسخی برای این پرسش هستیم که با توجه به اهداف و مبانی نظام اخلاقی اسلامی از منظر علامه مصباح یزدی (ره)[3] کدام معیار برای اخلاق هوش مصنوعی می‌تواند به عنوان الگوی مطلوب صورت‌بندی شود.

---

۱. با توجه به جایگاه «نیت» در دیدگاه اخلاقی علامه مصباح (ره) لازم به ذکر است که رویکردهای مختلفی به بحث امکان و چگونگی عاملیت اخلاقیِ هوش مصنوعی مطرح است اما با توجه به تمایزی که بین هوش مصنوعیِ قوی و هوش مصنوعی ضعیف وجود دارد و از سویی با توجه به اینکه هوش مصنوعی که در حال حاضر با کاربردهای وسیع در دسترس هست هنوز به سطح هوش مصنوعی قوی نرسیده است، مباحث عاملیت تام اخلاقی و به تبع آن بحث از نیت و قصدمندی مدنظر این پژوهش نیست (مصباح یزدی، ۱۳۹٤، نقد و بررسی مکاتب اخلاقی، ص ۳٤۰).

2. Ethics by Design

۳. محمدتقی مصباح یزدی (۱۳۱۳-۱۳۹۹)، فیلسوف، فقیه و اندیشمند معاصر ایرانی که از شاگردان برجسته علامه طباطبایی و امام خمینی به‌شمار می‌رفت و نقش مهمی در بازسازی فلسفه اخلاق اسلامی با رویکردی عقل‌گرایانه و الهیاتی داشت. آثار وی در تبیین اخلاق مبتنی بر غایت‌مندی الهی، نیت‌محوری و مطلق‌گرایی اخلاقی، جایگاه ویژه‌ای در میان پژوهشگران مسلمان داردو

۲. ضرورت و پیشینه مبانی ارزشی هوش مصنوعی

در دوران مختلف مکاتب فلسفی، تحت مقوله عقلانیت، حقیقت و معنابخشی به حیات بشری، در تلاش‌اند منش خاصی را در راستای اهداف خاص اجتماعی توصیه کنند که معنابخش به حیات بشری باشد. بر این اساس، هر چه ملاک معنا و عقلانیت حیات تغییر می‌کند و هر اندازه تفسیر از حقیقت مختلف می‌شود، رویکردها در جوانب مختلف حیات بشری هم مختلف می‌گردد. با ظهور فلسفه‌های جدید و پیشرفت‌های دیگر در حیات انسان، عقلانیت ارزش‌شناختی نیز مفهوم متعددی یافته و مبانی و اهداف متفاوتی را دنبال کرده است (جمشیدیها و نادی، ۱۳۹۹، ص ۴۷). در عصر حاضر، همسو با رشد فناوری‌های جدید، بویژه هوش مصنوعی، معنای عقلانیت ارزش‌شناختی توسعهٔ مفهومی فوق‌العاده داشته و جای خود را به عقلانیت ابزاری داده است. از آنجا که مبانی صاحبانِ سرمایه و تکنولوژی‌های پیشرفته، خاستگاهی اومانیستی دارد و فضای هستی‌شناختی و معرفت‌شناختی مدرن را ماتریالیسم و سکولاریسم فراگرفته است، عقل مدرن از عقل نظری، ارزشی و هدف‌گرا فاصله گرفته و به عقلانیت ابزاری، راهبردی و سودگرا بسنده کرده است. به همین سبب آرمان‌های معنوی و اخلاقی که جهان‌شمول و ارزش‌آفرین است و جامعه بشری را به فضیلت و تعالی معنوی برمی‌انگیزاند، جای خود را به اهداف مادّی، مقطعی و گذرا داده است و این امر دنیای مدرن را دچار بحران معنویت و اخلاق کرده است (ربانی گلپایگانی، ۱۳۸۳، ص ۵۴).

با گسترش ابزارهای تکنولوژیک و ورود این ابزارها به جوامع مختلف، این بحران معنویت و اخلاق به سرتاسر جهان صادر خواهد شد و مبانی و فرهنگ خود را بر ملت‌های مصرف‌کننده تحمیل خواهد کرد؛ به طوری که کوشش‌های فیلسوفان و مصلحان اجتماعی برای مهار بحران‌های اخلاقی کارساز نخواهد بود. بر همین اساس، لازم است که هر جامعه‌ای در مواجهه با تکنولوژی و ابزارهای پیشرفته، ضمن استفاده صحیح و کارآمد از آن، مبانی هستی‌شناختی، معرفت‌شناختی و ارزش‌شناختی آن را متناسب با فرهنگ و ارزش‌های خود بومی‌سازی کند.

بنابراین، در جهت توانمندسازی هوش مصنوعی برای استدلال در برابر مسائل اخلاقی، لازم است جوامع مختلف درطراحی و ساخت ماشین‌های هوشمند الگوریتم‌هایی را پیاده‌سازی کنند که تعاملات این ماشین‌ها با عامل‌های انسانی، متناسب با عقلانیت ارزش‌شناختی یا حکمت عملی توصیه شده در مکاتب فلسفی آن جامعه باشد. دستیابی به حکمت عملی و ارزش‌های اخلاقی هوش مصنوعی ضروری است؛ زیرا برای افزایش اعتماد کاربران به سیستم‌های هوش مصنوعی و

ربات‌ها برخورداری این سیستم‌ها از صلاحیت اخلاقی امری اجتناب‌ناپذیر است.

سامانه‌های هوش مصنوعی باید به گونه‌ای طراحی شوند که در فرایند تصمیم‌گیری، ارزش‌یابی و نیز اولویت‌های مربوط به ملاحظات اخلاقی، ذینفعان مختلف را در زمینه‌های چندفرهنگی مختلف بسنجد، استدلال خود را توضیح و شفافیت را تضمین کند. افزون بر آن مسئولیت انسانی طراحان و توسعه‌دهندگان برای گسترش سیستم‌های هوشمند در امتداد اصول و ارزش‌های اساسی انسانی در جهت اطمینان از شکوفایی و رفاه انسان‌ها در یک چارچوب اخلاقی مطلوب ضرورت دارد.

بر خلاف بسیاری از مکاتب که دایرهٔ ارزش‌های اخلاقی را محدود به محیط اجتماعی یا در نهایت به رابطه با خدا می‌دانند، در اسلام تمامی ارتباطات مناسب و مفید در جهت دستیابی به سعادت، مورد نظر قرار گرفته‌اند. بر این اساس، ارتباط انسان با خدا، با خود، خانواده، طبیعت و جامعه ارزش‌های ثابت و معینی دارند، یعنی هیچ مسئله‌ای از مسائل زندگی انسان نیست که تحت پوشش ارزش‌های اخلاقی اسلام قرار نگیرد (مصباح یزدی، ۱۳۹۴ الف، ص ۳۵۲). بنابراین، مبانی اخلاق اسلامی ظرفیت مناسبی برای صورت‌بندی الگوی مطلوب در نسبت اخلاق با هوش مصنوعی را در مواجه با رویکردهای رقیب در این حوزه داراست و می‌توان با تکیه بر نظریه فلاسفه و اخلاق‌پژوهان اسلامی این الگو را استخراج کرد.

در دههٔ اخیر، ضرورت‌ها و چالش‌های پیش‌روی فناوری‌ها سبب اقبال پژوهشگران و صاحب‌نظران به موضوعات مبنایی و بنیادین فناوری‌ها، بخصوص هوش مصنوعی گردیده و آثار علمی متعددی در این زمینه نگاشته شده است. از بین آثار غربی می‌توان به اختصار به برخی از مهم‌ترین آن‌ها اشاره کرد: ۱. کتاب اخلاق ربات: پیامدهای اخلاقی و اجتماعی رباتیک[1] به‌طور گسترده به موضوعات اخلاقی مرتبط با طراحی ربات‌ها می‌پردازد. این کتاب مجموعه‌ای از مقالات است که توسط متخصصان حوزه‌های مختلف نوشته شده و به بررسی چالش‌های اخلاقی در طراحی و استفاده از ربات‌ها می‌پردازد. از جمله یکی از مقالات مهم این کتاب در زمینه پیاده‌سازی اخلاق، مقاله‌ای است با عنوان «رویکرد فرمان الهی به اخلاق ربات‌ها»[2] نوشته «سلمر برینگزورد» و «جاشوا تیلور» است. در این مقاله استدلال شده است که ربات‌ها، به‌ویژه در زمینه‌های حسّاس مانند میدان جنگ، باید تحت نظارت اخلاقی دقیق قرار گیرند. آنها پیشنهاد

1. *Robot Ethics: The Ethical and Social Implications of Robotics*
2. *The Divine-Command Approach to Robot Ethics*

می‌کنند که اقـدامات مهمـی کـه توسـط چنـین ربات‌هـایی انجـام می‌شـود، بایـد از طریـق یـک چارچوب اخلاقی مبتنی بر فرمان‌های الهی تنظیم شود. ۲. کتـاب ماشـین‌های اخلاقـی: آمـوزش درست و نادرست به ربات‌ها[1] یکی از کتاب‌هـای بنیـادی در زمینـه اخـلاق محاسـباتی و طراحـی ماشین‌های اخلاقی است. این کتاب به بررسی ضرورت و چگونگی توسعه ربات‌هـایی می‌پـردازد که قادر به تصمیم‌گیری اخلاقی هستند. نویسندگان با معرفی مفهوم «اخلاق عملکردی»[2] بـر ایـن نکته تأکید می‌کنند که حتی اگر ربات‌ها به سـطح کامـل عاملیـت اخلاقـی نرسـیده باشـند، بایـد بتوانند رفتارهای خود را بر اساس اصول اخلاقی تنظیم کنند. ایـن کتـاب بـه دو رویکـرد اصـلی در طراحی می‌پردازد: ۱. رویکرد بالا بـه پـایین[3] کـه در آن، نظریـه‌های اخلاقـی ماننـد فایـده‌گرایی یـا وظیفه‌گرایی به صورت مستقیم در الگوریتم‌های ربات‌ها پیاده‌سـازی می‌شـوند. ۲. رویکـرد پـایین به بالا[4] که در آن، ربات‌ها مشابه فرایند رشد اخلاقـی در انسـان‌ها، از طریـق یـادگیری و تعامـل بـا محیط، اصول اخلاقی را به‌تدریج فرامی‌گیرند.

۳. مقالهٔ «پیاده‌سازی‌ها در اخلاق ماشین»[5] نوشته «سوزان تـالمیجر» و جمعـی از نویسـندگان به رویکردهای اصلی اخلاق در غرب اشاره شده و سه چـارچوب اخلاقـی فایـده‌گرا، وظیفـه‌گـرا و فضیلت‌گرا پرداخته شده است.

در ایران نیز می‌توان به سه اثر اشاره کرد: ۱.کتاب هـوش سـفید؛ از اخـلاق ماشـینی تـا ماشـین اخلاقی، نوشته جمعی از محقّقان است که در سال ۱۴۰۰ به چاپ رسیده است. در این کتـاب بـه مسائل فرااخلاق و اخلاق هنجاری با رویکردهای موجود در غرب پرداختـه اسـت؛ ۲. پایـان‌نامـه کارشناسی ارشد با عنوان «طراحی و پیاده‌سازی بعد اخلاقی در سیستم‌های چنـد عـاملی» نوشـته «میثم آزاد منجیری» با راهنمایی «ناصر قاسم آبادی» که در مؤسسه آمـوزش عـالی کـار قـزوین در سـال ۱۳۸۹ دفـاع شـده اسـت. ایـن تحقیـق بـا رویکـردی فنـی- مهندسـی بـه بعـد اخلاقـی هوش مصنوعی پرداخته است؛ ۳. از تازه‌ترین آثار در این زمینه نیز کتاب هوش مصـنوعی: فلسـفه، اخلاق، جامعه اثر «ابوطالب صفدری» است که در آن به جمع‌آوری و ترجمـه مقـالات خـارجی در پنج فصل پرداخته شده است. افزون بر این، مقالاتی نیز در زمینـه اخـلاق هـوش مصـنوعی بـه

1. *Moral Machines: Teaching Robots Right from Wrong*
2. Functional Morality
3. Top-Down
4. Bottom-Up
5. Implementations in Machine Ethics

رشته تحریر درآمده که همگی بر اخلاق کاربردی تکیه دارند.

تمایز پژوهش حاضر با آثاری که اشاره شد، این است که در اینجا تلاش شده است ناظر به مرحلهٔ طراحی و پیاده‌سازی نظریه‌های اخلاقی در هوش مصنوعی با توجه به اخلاق اسلامی و با تکیه بر مبانی اخلاقی علامه مصباح (ره) الگوی مطلوب را استخراج و ارائه دهد تا زمینه مناسبی در طراحی الگوریتم‌های هوش مصنوعی متناسب با فرهنگ بومی خود ارائه دهد.

۳. طبقه‌بندی نسبت اخلاق با هوش مصنوعی

نگرانی‌های اخلاقی در مورد هوش مصنوعی را می‌توان در دو سطح دسته‌بندی کرد: ۱) نگرانی‌های اخلاقی ناشی از طراحی هوش مصنوعی؛ ۲) نگرانی‌های اخلاقی که از تعاملات انسان و هوش مصنوعی ناشی می‌شوند (Giarmoleo et al, 2024, p. 1). آنچه در این مقاله مدِّ نظر است بررسی سطوح نسبت اخلاق با هوش مصنوعی بر اساس نگرانی‌هایی است که از طراحی هوش مصنوعی ناشی می‌شود. هرچند این نگرانی‌ها در مُدل‌هایاولیه هوش مصنوعی کمتر وجود داشت، اما با پیشرفت‌های جدید و پیچیده‌تر شدن الگوریتم‌های هوش مصنوعی این نگرانی‌ها به سطح بی‌سابقه‌ای در مدل‌های سیستم‌های خودمختار هوش مصنوعی رسیده است. به اذعان فناوران هوش مصنوعی «برای به‌دست آوردن مزیّت‌های بالقوه سیستم‌های هوشمند مستقل (خودمختار)، طراحی و توسعه آنها باید با ارزش‌های اساسی و اصول اخلاقی همسو باشد» (Leikas, 2019, p. 1).

نسبت اخلاق با هوش مصنوعی مبتنی بر فهمی که از چیستی و غایت اخلاق از یک‌سو و ماهیت و هدف هوش مصنوعی از سوی دیگر وجود دارد، می‌تواند دارای صورت‌بندی‌های متفاوتی در بخش طراحی باشد. ملاحظات فوق سبب شده تا اخلاق‌پژوهان هوش مصنوعی با هدف ایجاد امنیت، اعتماد و مسئولیت‌پذیری این فناوری، رابطهٔ اخلاق با هوش مصنوعی را در سه سطح طبقه‌بندی کنند (Dignum, 2018, p. 2):

- اخلاق در طراحی[1] (یا اخلاق برای طراحان)؛
- اخلاق برای طراحی[2] (یا اخلاق برای کاربران)؛
- اخلاق با طراحی[3] (یا اخلاق آمیخته با طراحی)؛

1. Ethics in Design
2. Ethics for Design
3. Ethics by Design

۳. ۱. اخلاق در طراحی

اخلاق در طراحی، به مجموعه روش‌های نظارتی و مهندسی اشاره دارد که از تجزیه و تحلیل و ارزیابی مفاهیم اخلاقی در سیستم‌های هوش مصنوعی پشتیبانی می‌کنند (Dignum, 2018, p. 2)؛ به‌ویژه زمانی که این سیستم‌ها جایگزین یا مکمّل ساختارهای اجتماعی سنّتی می‌شوند. از این سطح از رابطه میان اخلاق و هوش مصنوعی می‌توان به «اخلاق برای طراحان و توسعه‌دهندگان هوش مصنوعی» نیز تعبیر کرد؛ زیرا طراحی را مبتنی بر رعایت اصول اخلاقی جامع و پیش‌فرض‌های ارزشی مقبول ممکن می‌سازد؛ تا از این طریق بر اعتماد کاربران جهت بهره‌گیری از این فناوری بیافزاید. با رعایت چارچوب‌های تعیین شده در این سطح، بخشی از نگرانی‌ها پیرامون مسئولیت اخلاقی تصمیمات اخذ شده توسط سیستم‌های خودمختار کاسته خواهد شد.

۳. ۲. اخلاق برای طراحی

این سطح از اخلاق ناظر بر کاربران است و شامل آئین‌نامه‌های رفتاری، استانداردها و فرایندهای صدور گواهینامه‌ای است که برای تضمین یکپارچگی توسعه‌دهندگان و کاربران در مراحل تحقیق، طراحی، ساخت، استفاده و مدیریت سیستم‌های هوشمند تدوین می‌شود (Dignum, 2018, p. 2).

هرچند که دستیابی به اهداف ارزشی در استفاده از ابزارهای هوشمند در نتیجهٔ بکارگیری اصول اخلاقی صحیح در هر سه سطح است، اما با توجه به بدیع و نوظهور بودن تصمیمات اخلاق در ماشین‌های خودمختار، آنچه امروزه به عنوان خطرات احتمالی هوش مصنوعی بیشتر مورد توجه اخلاق پژوهان هوش مصنوعی است، سطح اخلاق با طراحی است که در این پژوهش نیز مورد تأکید است.

۳. ۳. اخلاق با طراحی (کاربست اخلاق در ماشین)

هوش مصنوعی (AI) به سرعت به بخشی جدایی‌ناپذیر از فرایندهای تصمیم‌گیری در صنایع مختلف تبدیل شده است و ادغام آن در تصمیم‌گیری، نگرانی‌های اخلاقی بی‌شماری را بوجود آورده است که نیازمند بررسی دقیق است. ملاحظات فوق سبب شده تا برخی از اخلاق‌پژوهان هوش مصنوعی با هدف ایجاد امنیت، اعتماد و مسئولیت‌پذیری این فناوری، مسألهٔ اخلاق با طراحی را مطرح نمایند (Dignum, 2018, p. 2).

اخلاق با طراحی (اخلاق آمیخته با طراحی) با هدف «ادغام فنی الگوریتمی قابلیت‌های استدلال اخلاقی به عنوان بخشی از رفتار سیستم مستقل به گونه‌ای تنظیم می‌شود که سیستم قادر باشد یا خودش تصمیمات اخلاقی بگیرد یا به کاربران هشدار دهد یا بر انحرافات احتمالی رفتار از اصول اخلاقی نظارت کند» (Dignum, 2018, pp. 2-3). این سطح از اخلاق به دنبال آن است تا حدِّ امکان از نگرانی‌های ناشی از تصمیم‌سازی‌های مستقل ماشین بکاهد؛ بنابراین، هدف آن، گنجاندن (پیاده سازی) سیستماتیک و جامع ملاحظات اخلاقی در فرایند طراحی و توسعه سیستم‌های هوش مصنوعی است (Brey & Dainow, 2023, p. 1). به عنوان نمونه، وجود سوگیری در الگوریتم‌های هوش مصنوعی زمینه‌ساز چالش‌های اخلاقی مهمی است. مدل‌های یادگیری ماشینی از داده‌های تاریخی یاد می‌گیرند و اگر این داده‌ها مغرضانه باشند، سیستم هوش مصنوعی ممکن است سوگیری‌های موجود را تداوم بخشد و حتی تشدید کند. پرداختن به چنین پتانسیل‌های ضداخلاقی مستلزم بررسی دقیق داده‌های آموزشی، طراحی الگوریتمی فاقد سوگیری و نظارت مداوم برای اطمینان از انصاف و کاهش تبعیض است (Osasona et al, 2024, p. 322).

۴. رویکردهای معرفت اخلاق در انسان

از آنجا که هدف در اخلاق با طراحی توانمندسازی سیستم‌های هوشمند مستقل جهت رعایت ملاحظات و ارزش‌های اخلاقی در تصمیم‌سازی ماشینی است، پیش از اشاره به چگونگی توانمندسازی و پیاده‌سازی آن، لازم است به این پرسش بنیادین پاسخ داد که معرفت اخلاقی در انسان چگونه شکل می‌گیرد؟ آیا اخلاقی یا غیراخلاقی بودن یک عمل تابع استدلال است یا تنها بر اساس شهود و تجربهٔ زیستهٔ فردی است؟ به عبارت دیگر، آیا اصول از پیش تعیین شده‌ای دربارهٔ اخلاقی بودن اعمال وجود دارد یا خیر؟ بایدها و نبایدهای اخلاقی مبتنی بر چیست؟ نخستین طبقه‌بندی پیرامون پرسش‌های یادشده در سال ۲۰۰۵ توسط آلن و همکارانش انجام گرفت (Allen et al, 2005, p. 152). در این طبقه‌بندی سه دیدگاه غالب «اصل‌گرایی»[1]، «نمونه‌گرایی»[2] و «تعادل تأملی»[3] مطرح شده است.

بر اساس «اصل‌گرایی»، رفتاری اخلاقی یا غیراخلاقی محسوب می‌شود که اصولی از پیش

1. Generalism
2. Particularism
3. Reflective Equilibrium

تعیین شده برای آن وجود داشته باشد. استدلال‌های اخلاقی از اصول کلی شروع می‌شوند و به مصادیق هر عمل تسرّی می‌یابند. اصولی مانند «دروغ نگو» و «به امانت وفادار باش» از این قسم هستند. اکثر نظریه‌های اخلاق هنجاری از جمله «وظیفه‌گرایی» و «فایده‌گرایی» چنین رویکردی دارند. در مقابل «اصل‌گرایی»، «نمونه‌گرایی» با دو خوانش اِفراطی و معتدل مطرح است.

بر اساس خوانش اِفراطی «نمونه‌گرایی»، نه‌تنها وجود اصول کلی اخلاقی ضرورتی ندارد، بلکه حتی التزام به اصول اخلاقی مانع از انجام عمل درست می‌شود و رفتار اخلاقی بر اساس شرایط و موقعیت‌ها شکل می‌گیرد. در خوانش معتدل «نمونه‌گرایی» وجود اصول اخلاقی قابل انکار نیست، اما نه به صورت پیشینی، بلکه این اصول از نمونه‌های عینی و شهودات افراد استخراج می‌شوند. بدین صورت که ابتدا حکم اخلاقی از نمونه‌ها و موقعیت‌ها استخراج می‌گردد و در نهایت، با بررسی همین نمونه‌ها به استخراج اصول اخلاقی منتهی می‌شود. بنابراین، وجه اشتراک هر دو خوانش آن است که این موقعیت‌ها، مصادیق و نمونه‌های هر عمل هستند که اولویت دارند و نه اصول از پیش تعیین شده (عبدخدایی و دیگران، ۱۴۰۰، ص ۸۷-۸۸).

دیدگاه «تعادل تأملی» در صدد برقرار کردن تعادل میان دو دیدگاه سابق است. نظریهٔ اخلاقی «دکترین اثر مضاعف»[1] یکی از نمونه‌های شاخص دیدگاه «تعادل تأملی» است.[2] این نظریه، یک نظریهٔ اخلاقی است که برای تصمیم‌گیری در موقعیت‌های دشوار اخلاقی که در آنها اقدام همراه با تأثیرات مثبت و منفی غیرقابل اجتناب به نظر می‌رسند، به کار می‌رود.

در «دکترین اثر مضاعف» از یک‌سو به قواعد اخلاقی مشخصی مانند عدم قصد ضرر به دیگران و عدم استفاده از ضرر به عنوان وسیله پای‌بند است و از سوی دیگر، با شرط توازن میان فایده و ضرر، انعطاف‌پذیری لازم برای ایجاد تعادل در شرایط خاص را فراهم می‌کند (Govindarajulu & Bringsjord, 2017, p. 4722). به عبارتی، در این دیدگاه طی فرایندی می‌توان به تعادل میان اصول کلی و نمونه‌های کاربردی دست یافت. در این فرایند گاهی اصول به نفع نمونه‌ها و مصادیق اصلاح می‌گردند و گاهی مصادیق به نفع اصول کنار گذاشته می‌شوند و قضاوت‌های اخلاقی امری قطعی و بسته نیستند. اصول کلی اخلاقی راهنمای عمل بوده، اما همواره در جهت و

1. the Doctrine of Double Effect / DDE

۲. نظام‌های مطلق‌گرای اخلاق در چند جبهه مورد انتقاد شدید قرار گرفته‌اند. اصل اثر مضاعف توسط علمای اخلاق کاتولیک برای غلبه بر این مخالفت ها تدوین شد. طرفداران DDE ادعا دارند که این اصل برای اجتناب از مشکلات ناشی از رعایت مجموعه از قوانین اخلاقی مطلق، یک الزام ضروری است.

ناظر به موارد کاربردی عمل می‌کنند (عبدخدایی و دیگران، ۱۴۰۰، ص ۸۹). لازم بـه ذکـر اسـت کـه ایـن نظریه اخلاقی تمایز جدّی با نظریه علامه مصباح (ره) دارد که در ادامه، خواهد آمد.

۵. شیوه‌های متناظر در اخلاق با طراحی

اخلاق با طراحی به عنوان یک رویکرد جامع به دنبال ادغام اصـول اخلاقـی بـه طـور سیسـتماتیک در فرایند طراحی و توسعه سیستم‌های هوش مصنوعی اسـت. در ایـن زمینـه، سـه شـیوهٔ «بـالا بـه پایین» و «پایین به بالا» و «تلفیقی» می‌توانند به عنوان شیوه‌های مختلف برای پیاده‌سـازی اخـلاق در طراحی سیستم‌های هوش مصنوعی عمل کنند.

۵. ۱. شیوهٔ «پایین به بالا»

در رویکرد «پایین به بالا»، به جای پایبندی به اصول کلی، ماشین در مشابهت با ذهـن انسـان طـی یک فرایند تکاملی یا یادگیری، با توجه به شرایطی کـه بـا آن در تعامـل اسـت اسـتدلال اخلاقـی را کسب می‌کند. این شیوه تلاشی برای آموزش و تکامـل عامل‌هـایی اسـت کـه رفتـار آنهـا تقلیـد از رفتار اخلاقی شایسته انسانی است (Allen et al, 2005, p. 149). برای مثـال، اگـر فـرض شـود ربـاتی برای مراقبت از کودکان در مهد کودک طراحی شده است. او ابتدا هیچ اصل اخلاقی را نمـی‌دانـد، اما از طریق یادگیری تقویتی و مشاهده رفتـار مربیـان انسـانی مـی‌آمـوزد کـه هنگـام زمین‌خـوردن کودکان باید به آن‌ها کمک کرد. از این‌رو، کمک کـردن در شـرایط خـاص را بـه عنـوان یـک اصـل اخلاقی فرامی‌گیرد و در کنش‌های بعدی رعایت می‌کند.

در این روش، نظریه اخلاقی خاصـی را تحمیـل نمی‌کننـد، بلکـه بـه جـای آن، محیط‌هـایی را فراهم می‌آورند که در آن رفتارهای مناسب انتخاب شده یـا پـاداش داده می‌شـوند. ایـن رویکردهـا شامل یادگیری تدریجی از طریق تجربه است؛ خواه بـه صـورت مکـانیکی و ناخودآگـاه از طریـق آزمون و خطای تکاملی، خواه از طریق دست‌کاری برنامه‌نویسان یـا مهندسـان هنگـام مواجهـه بـا چالش‌های جدید، یا از طریق توسعه آموزشـی یـک ماشـین یادگیرنـده. هـر یـک از ایـن روش‌هـا شباهت‌هایی با شیوه‌ای دارند که یک کودک خردسـال در یـک بافـت اجتمـاعی آمـوزش اخلاقـی دریافت می‌کند که در آن رفتارهای مناسب و نامناسب شناسـایی می‌شـوند، بـدون آن‌کـه نظریـه‌ای صریح دربارهٔ معیارهای این رفتارها ارائه شود (Allen et al, 2005, p. 151).

در رویکرد پایین به بالا افزون بر توجه به نقش یادگیری در رشد اخلاقی، تأثیرات طبیعت و ژنتیک در تکامل اخلاقی نیز مدِّ نظر است. بر این اساس، در این رویکرد دو نظریه «تکامل‌محور»[1] و «یادگیری محور»[2] مطرح می‌شود (عبدخدایی و دیگران، ۱۴۰۰، ص ۱۰۶-۱۰۷). بنابراین، بر خلاف نظریه‌های اخلاقی از بالا به پایین که اخلاقی بودن و نبودن را تعریف می‌کنند، در رویکردهای پایین به بالا، اصول اخلاقی، در مکاتب واقع‌گرا، باید کشف یا، در مکاتب غیرواقع‌گرا، بایستی ساخته شود (Wallach & Allen, 2009, p. 80). بر این اساس، این روش با رویکرد نمونه‌گرایی در اخلاق سازگاری بیشتری دارد و امکان پیاده‌سازی این رویکرد اخلاقی در ماشین را فراهم می‌سازد.

۵. ۲. شیوهٔ «تلفیقی»

در شیوهٔ «تلفیقی»، اصول بالا به پایین نقش مهمی در هدایت رفتارِ عاملِ اخلاقیِ هوشمند دارند. از طرف دیگر، رویکرد پایین به بالا ضامن انعطاف‌پذیری در رفتار عاملِ اخلاقیِ هوشمند هستند؛ به طوری که سبب خواهند شد تا واکنش عاملِ اخلاقی فراتر از پیروی محض از اصول و قواعد از پیش تعیین‌شده باشد (عبدخدایی و دیگران، ۱۴۰۰، ص ۱۱۴). به عنوان مثال، ربات دستیار پزشکی با پایبندی به اصل رازداری، اطلاعات حسّاس بیمار را محرمانه تلقی می‌کند، اما با توجه به تجربه‌های یادگرفته‌شده، در شرایط بحرانی تصمیم می‌گیرد برای حفظ جان بیمار، اطلاعات را محدوداً با پزشک در میان بگذارد. این تصمیم تلفیقی از اصول ثابت اخلاقی (بالا به پایین) و یادگیری موقعیتی (پایین به بالا) است. به این ترتیب رویکرد تلفیقی، هم پایبندی به اصول را حفظ می‌کند و هم انعطاف‌پذیری را تضمین می‌نماید. این شیوه با رویکرد تعادل تأملی در اخلاق قابل تطبیق است و«فضیلت‌گرایی» و «دکترین اثر مضاعف» دو نظریه غالب در این رویکرد هستند.[3] هر سه رویکرد فوق برآن‌اند تا بر اساس فرایند شکل‌گیری معرفت اخلاقی در انسان، کاربست اخلاق با طراحی را مورد بررسی قرار دهند.

1. Evolutionary-inspired Approaches
2. Learning-based Approaches

۳. هر یک از این نظریات به فراخور مبانی و اصولی که دنبال می‌کنند، با چالش هایی مواجه هستند که مهندسان و فلاسفه معاصر سعی در به حداقل رساندن این چالش ها دارند.

۵. ۳. شیوه «بالا به پایین»[1]

در شیوه «بالا به پایین» مجموعه‌ای از قواعد و اصول کلی اخلاقی که قابلیت تبدیل بـه الگـوریتم را دارند؛ مبنای اخلاق ماشین قرار می‌گیرند. در این روش، سعی بر آن است کـه اصـول و نظریـه‌های روشن اخلاقـی بـه الگوریتم‌هـا تبـدیل شـوند (Allen et al, 2005, p. 149). بـر ایـن اسـاس، اصـول اخلاقی از پیش تعیین‌شده در ماشین پیاده‌سازی‌شده و سپس ماشـین در موقعیت‌هـای مختلـف آن را اعمال مـی‌کنـد (عبدخـدایی و دیگـران، ۱۴۰۰، ص ۹۰). بـه عنـوان مثـال، ربـاتی کـه بـرای مراقبـت از سالمندان طراحی شده است، به گونه‌ای برنامه‌نویسی و طراحی‌شده است که بر اساس اصـل کلـی «همیشه حقیقت را بگو» حق ندارد دربارهٔ وضعیت جسمی یا بیماری احتمـالی او دروغ بگویـد تـا برای مثال، حال او را از لحاظ روانی بهتر جلوه دهد.

از آنجا که شیوه «بـالا بـه پـایین» بـه دنبـال تضـمین تطـابق سیسـتم‌های هـوش مصـنوعی بـا استانداردهای اخلاقی از پیش تعریف‌شده اسـت و مسـئولیت سیسـتم‌ها بـرای اتخـاذ تصـمیمات اخلاقی بر اساس قواعد و اصول خاص تعیین شده، ایـن روش منطبـق بـر مـدل‌های قاعـده‌محور[2] در تصـمیم‌گیری در هـوش مصـنوعی خواهـد بـود کـه اغلـب از رویکـرد منطـق نمـادین[3] در هوش مصنوعی بهره می‌گیرد و در حوزه‌هایی که شفافیت و تضـمین اخلاقـی اهمیـت دارد، ماننـد پزشکی یا حقوق، مورد استفاده قرار می‌گیرد. بـه نظـر مـی‌رسـد روش «بـالا بـه پـایین» مـی‌توانـد همسو با رویکرد اصل‌گرایی تلقی شود.

بر این اساس، اگر بتوان اصول یا قواعد اخلاقی را به صراحت بیان کرد، رفتار اخلاقی صرفاً تابع قـوانین خواهد شد و عامل‌های اخلاقی هوشمند تنها با محاسبه دقیق قادر خواهند بـود تصـمیم بگیرنـد کـه آیـا اقدامات آنها اخلاقی است یا خیر[4] (Wallach & Allen, 2009, p. 83). به اعتقاد صـاحب‌نظـران در میـان مکاتب فلسفی اصل‌گرا، دو مکتب «فایده‌گرایی» و «وظیفـه‌گرایـی» قابلیـت تبـدیل بـه الگـوریتم و پیاده‌سازی در ماشین را بـا تکیـه بـر همـین روش، دارنـد (Wallach & Allen, 2009, p. 85). هرچنـد ایـن

1. Top-down
2. Rule-Based Models
3. Symbolic Logic

٤. لازم به ذکر است با توجه به جایگاه و نقش نیت در ارزش عمل اخلاقـی در دیـدگاه اخلاقـی اسـتاد مصـباح (ره) ایـن مسئله مطرح خواهد شد آیا ماشین می‌تواند نیت اخلاقی داشته باشد تا عمل او اخلاقی تلقی شده و او به عنوان عامـل اخلاقی شناخته شود؟ این مسئله ناظر به بخش فرااخلاق و هوش مصنوعی است و حال آنکه مقاله حاضر در بخـش اخلاق هنجاری و هوش مصنوعی نگارش یافته است. از این رو نیازمند تحقیقی مستقل است.

شیوه، در بدو امر، مُدلی مطلوب برای پیاده‌سازی اخلاق با طراحی به نظر می‌رسید، اما با چالش‌هایی از جمله عدم درک مفاهیم موجود در قواعد و تعمیم آنها مواجه شد و این امر سبب شد تا شیوهٔ دیگری معرفی شود (عبدخدایی و دیگران، ۱۴۰۰، ص ۱۰۲).

۶. نظریهٔ اخلاقی علامه مصباح یزدی

همان‌طور که بیان شد، مسئلهٔ اصلی این پژوهش ارائه صورت‌بندی نسبت اخلاق با هوش مصنوعی از منظر علامه مصباح در مواجهه با رویکردهای دیگر است. برای این منظور پس از ذکر مطالبی که بیان گردید، ابتدا لازم است مشخص شود که دیدگاه اخلاقی ایشان چه نسبتی با دیدگاه‌های سه‌گانه اصل‌گرا، نمونه‌گرا یا تعادل تأملی دارد و سپس اگر بنا باشد این دیدگاه اخلاقی در سطح اخلاق با طراحی مطالعه شود، ذیل کدام رویکرد قابلیت پیاده‌سازی را خواهد داشت.

علامه مصباح ـ در اخلاق هنجاری ـ تبیینی غایت‌گرایانه از دیدگاه اخلاقی اسلام ارائه می‌دهد. ایشان با نقد و بررسی دیدگاه فضیلت‌گرایی غایت‌گرایانه یونانیان و با تأکید بر چهارده اصل لازم در تبیین نظام اخلاقی اسلام، ارزش اخلاقی فعل اختیاری انسان را تابع تأثیر آن فعل در رسیدن انسان به کمال حقیقی انسانی می‌داند (مصباح یزدی، الف ۱۳۹۴، ص ۳۳۹) و با تبیین شش عنصر و مولفه اصلی در نظریه اخلاقی اسلام، کمال نهایی انسان را قُرب‌الهی معرفی می‌کند (مصباح یزدی، الف ۱۳۹۴، ص ۳۴۳). بر این اساس، می‌توان دیدگاه اخلاقی ایشان را «غایت‌گرایی قُرب‌الهی» نامید. از این‌رو، مبنای نظام اخلاقی ایشان از سویی هم بر اصول و معیارهای از پیش‌تعیین‌شده مبتنی است و هم افزون بر تبیین اصول کلی اخلاقی به معیارهای ویژهٔ فهم مصادیق و نمونه‌های افعال اختیاری نیز التفات دارند.

علامه مصباح (ره) در تبیین ارتباط اخلاق و اسلام، عقل را در تحلیل کمال نهایی و استخراج مفاهیم کلی که رابطهٔ میان افعال اختیاری و کمال نهایی را تبیین می‌کند، کارآمد می‌داند، اما بر این نکته تأکید دارد که این مفاهیم کلی در تعیین مصادیق دستورهای اخلاقی چندان کارایی ندارند؛ زیرا عقل به تنهایی قادر به تشخیص رابطه افعال اختیاری با غایات آنها و تأثیرات دنیوی و اخروی‌شان نیست. پس برای اینکه مصادیق خاص دستورهای اخلاقی را به دست آوریم، نیازمند دین هستیم. این وحی است که دستورهای اخلاقی را در هر مورد خاصی با حدود خاص خودش و با شرایط و لوازمش تبیین می‌کند و عقل به‌تنهایی از عهدهٔ چنین کاری برنمی‌آید (مصباح یزدی، ب ۱۳۹۴، ص ۲۳۳).

در این رویکرد، افزون بر اصول کلی که به واسطهٔ عقلِ عملی کشف می‌شوند، مصادیق و نمونه‌های کاربردی اخلاق از طریق وحی تبیین می‌شوند. از این‌رو، می‌توان در نسبت با سه‌گانه‌ای که پیش‌تر بیان گردید، نظریه علامه مصباح (ره) را ذیل دیدگاه تعادل تأملی جای داد، هرچند لازم است به تمایزات دیدگاه ایشان با دیدگاه تعادل تأملی که خاستگاه آن در مکاتب غربی است، ملتزم بود.

در حقیقت، روش تعادل تأملی طریقی بینابینی میان اصل‌گرایی و نمونه‌گرایی است که در آن، میان اصول کلی و مصادیق جزئی تعادل برقرار می‌شود. در این روش، اصول کلی اخلاقی می‌توانند خاستگاه‌های متفاوتی داشته باشند. این اصول می‌توانند برگرفته از دین، فرهنگ، ادبیات، فلسفه و نظایر آن باشند. قاعده حد وسط در فضیلت‌گرایی و سه اصل اثر مضاعف در «دکترین اثر مضاعف» (Govindarajulu & Bringsjord, 2017, p. 4722) نمونه‌هایی از اصول و قواعد کلی حاکم بر مکاتب اخلاقی تعادل تأملی هستند؛ امادر حوزهٔ مصادیق کاربردی، این مصادیق بر هیچ اصلی پیشینی مبتنی نیستند، بلکه آنچه را که «خوب» است توسط خود شخص در شرایط و موقعیت‌های مختلف استخراج می‌گردد، از این‌رو، با نوعی نسبیت اخلاقی مواجهیم. بنابراین، در قرائت غربی از رویکرد تعادل تأملی که ریشه در مکاتب اومانیستی دارد، ملاک تشخیص افعال اختیاری «خوب»، خود انسان بسته به شرایط و اقتضائات خاص است. در حالی که در نظریهٔ علامه مصباح یزدی اصول کلی توسط عقل اتخاذ می‌گردد، اما ملاک تشخیص مصداق فعل اخلاقی «خوب» وحی است. از این‌رو، در نظریهٔ اخلاقی، ایشان ضمن التفات به محوریت «توحید» وحی تبیین‌گر ارزش اخلاقی افعال است.

در ادامه، ضمن مقایسه دیدگاه فضیلت‌گرایی با نظریه اخلاقی علامه مصباح ذیل رویکرد تلفیقی در کاربست اخلاق با طراحی، نشان داده شده که چگونه می‌توان از آموزه‌های ایشان در شکل‌دهی الگوی اخلاقی در هوش مصنوعی و فناوری‌های نوین بهره برد.

۶. ۱. تبیین دیدگاه علامه مصباح در قیاس با دیدگاه فضیلت‌گرایی

در رویکرد تلفیقی از نظریهٔ اخلاق فضیلت‌گرا استفاده می‌گردد. فضیلت‌گرایی به عنوان یکی از مهم‌ترین نظریه‌های اخلاقی از زمان ارسطو و افلاطون مطرح بوده است و در طول تاریخ تفاسیر متعددی از آن صورت گرفته است، اما در مجموع خود را رقیب نظریات غایت‌گرایانه و وظیفه‌گرایانه اخلاقی می‌داند. از منظر ارسطو، فضایل به دو دسته فضایل عقلانی و فضایل

اخلاقی تقسیم می‌گردند. فضایل عقلانی در بردارندهٔ اصول و قواعد کلی اخلاقی است که از نظر فلاسفه علم معاصر، به عنوان باورهای پایه یا سنگ‌های صلب اعتقادی، ارزشی و اخلاقی شناخته شده و تحت تأثیر بسترهای اجتماعی و فرهنگی قرار نمی‌گیرند. این ارزش‌ها در جوامع گوناگون مشترک بوده و معمولاً یا تغییر نمی‌کنند و یا به سختی قابل تغییر هستند. این باورها بنیان تفکرات هر فرد و جامعه را بنا می‌کنند. از این‌رو، رویکرد بالا به پائین برای پیاده‌سازی این نظریه‌ها و اصول در ماشین پیشنهاد می‌گردد. از سوی دیگر، برخی ارزش‌ها و باورهای اخلاقی که ارسطو از آنان به فضایل اخلاقی نام می‌برد، با تکرار و تمرین برای فرد حاصل می‌آید. به بیان فلاسفه، اخلاق وابسته به نرم‌ها و هنجارهای فرهنگی و اجتماعی هستند. این باورها نه‌تنها در جوامع گوناگون متفاوت هستند، بلکه تحت شرایط مختلف نیز تغییر خواهند کرد. بنابراین، نمی‌توان قاعده کلی برای آنها تدوین نمود، بلکه باید تحت شرایط گوناگون، عملکرد متفاوتی اتخاذ شود. از این‌رو، رویکرد یادگیری و تعمیم مصادیق یا به عبارتی، رویکرد پائین به بالا قابل توجه و تأمل است (عبدخدایی و دیگران، ۱۴۰۰، ص ۱۸۵-۱۸۶). به اعتقاد ارسطو، هیچ قاعده مشخصی برای رسیدن به ارزش‌های اخلاقی وجود ندارد، بلکه به واسطه بررسی موقعیت‌های مشخص توسط افراد بین وسیله و هدف ارتباط برقرار می‌گردد. بر این اساس، افراد عمل «خوب» را به واسطه شهود، تجربه و استقرا یاد می‌گیرند (Aristotle, 1140a-1141b). از این‌رو، در سیستم‌های هوشمند نیز عمل «خوب» در طول زمان و به واسطه یادگیری از داده‌ها تعیین می‌گردد.

در رویکرد بالا به پایین نظام اخلاقی علامه نیز عقل فارغ از وحی اصول کلی (مانند عدل خوب است) را اتخاذ می‌کند. در این اصول، هیچ اعتقادی اخذ نشده است و هر کس می‌تواند این اصل را بپذیرد، بی‌آنکه به اعتقادات یا دستورهای دینی پای‌بند باشد، اما در تعیین کمال نهایی و رابطه آن با افعال اختیاری به شدّت به اصول و اعتقادات دینی و محتوای وحی نیازمند است (مصباح یزدی، ۱۳۹۴ب، ص ۲۳۲)، اما در رویکرد پایین به بالا و در تعیین مصادیق و نمونه‌های کاربردی (مثل اینکه عدل در موارد مختلف چه اقتضائاتی دارد)، شهود افراد ملاک استخراج ارزش‌های اخلاقی نیست، بلکه دستورات اخلاقی در هر مورد خاص با حدود و شرایط و لوازمش به وسیله وحی تبیین می‌شوند. به عبارتی، در رویکرد پایین به بالای این نظریه، یادگیری اساس اصول اخلاقی نیست، بلکه وحی تعیین‌گر ارزش‌های اخلاقی است (مصباح یزدی، ۱۳۹۴ب، ص ۲۳۳).

محوریت «وحی» در رویکرد پایین به بالا منجر به تمایز این دیدگاه از دیگر رویکردهای «تعادل تأملی» همچون دکترین اثر مضاعف است؛ زیرا همان‌طور که در توصیف «دکترین اثر

مضاعف» گذشت، در این نظریه گاهی اصول به نفع نمونه‌ها و مصادیق اصلاح می‌گردند و گاهی مصادیق به نفع اصول کنار گذاشته می‌شوند و قضاوت‌های اخلاقی امری قطعی و بسته نیستند و از آنجا که نمونه‌ها و مصادیق جزئی نسبت به موقعیت‌های مختلف، متفاوت قضاوت می‌شوند، از این‌رو با وجود سه اصل کلی در آن، این نظریه منجر به نوعی نسبیت در اخلاق خواهد شد، اما پیش‌فرض در «غایت‌گرایی قُرب‌الهی» و به طور عام در همه مکاتب اخلاق اسلام این است که تعارضی میان عقل و دین نیست و گزاره‌های عقل‌ستیز در اسلام وجود ندارد و ، بلکه دین اسلام تنها شامل گزاره‌های عقل‌پذیر و عقل‌گریز است که در دستهٔ اول هرآنچه را که عقل سلیم می‌فهمد، شرع نیز تأیید می‌کند و در دستهٔ دوم آنچه را که وحی به آن حکم می‌کند، از توان درک عقل خارج است. بنابراین، در دیدگاه «تعادل تأملی» علامه مصباح یزدی (ره) نمی‌توان مصداقی را یافت که وحی به آن حکم کرده باشد، اما با اصول کلی اخلاقی که عقل کاشف از آن است، در تعارض باشد و همواره عقل به عنوان حجّت درون با وحی که حجّتی بیرونی است در توافق است و احکام اخلاقی آن قطعی و مطلق‌اند.

با این وجود، دیدگاه «غایت‌گرایی قُرب‌الهی» علامه مصباح یزدی از جمله رویکردهای «تعادل تأملی» است که با توجه به آن، در پیاده‌سازی الگوریتم‌های اخلاقی، بایستی به دنبال مُدلی کاربردی بود که بتواند الزامات اخلاقی این نگرش غایت‌گرایانه را تأمین کند.

۷. رویکرد «غایت‌گرایی زمینه‌ساز کمال»

هر یک از مکاتب فلسفهٔ اخلاق بر اصول موضوعهٔ خاصی مبتنی است. نظریهٔ «غایت‌گرایی قُرب‌الهی» نیز بر چهارده اصل استوار است (مصباح یزدی، ۱۳۹۴ الف، ص ۳۳۰ـ۳۳۱) که مهم‌ترین این اصول جهت‌گیری نفس انسان به وسیله فعل اختیاری به سوی خداوند و رسیدن به قُرب‌الهی است (مصباح یزدی، ۱۳۹۴ الف، ص ۳۳۶). نفس انسان با این که موجود واحد بسیطی است، موجودی ذومراتب بوده و دارای مراتب طولی و شئون عرضی است. توجه نفس، در یکی از مراتب به یکی از شئونش منجر به پیدایش فعل اختیاری خواهد شد. حال این توجه یا به خودی‌خود و از ناحیهٔ خود نفس است و یا در اثر محرک‌های خارجی و عوامل جاذب بیرونی است و موجب کمالی برای نفس در یک مرتبه و شأن خاص می‌گردد، اما این کمالات همیشه مربوط به مراتب عالی نفس نیستند، بلکه بسته به شأنی از شئون نفس، کمالی نسبی را برای نفس حاصل می‌کند (مصباح یزدی، ۱۳۹۴ الف، ص ۳۳۴). کمالات زمانی به کمال حقیقی نفس منتهی می‌شوند که بر جوهر آن در تمامیت و مراتب

گوناگونش اثر بگذارند. به بیان دیگر، اگر یک کمال نسبی در مرتبه‌ای خاص یا شأنی از شئون نـازل نفس تأثیر بگذارد، این تأثیر باید به‌گونه‌ای باشد که موجب تکامـل کلـی نفـس شـود؛ زیـرا نفـس، حقیقتی واحد و ذومراتب است که تمامی ابعاد آن در یک سِیر هماهنگ به کمال می‌رسـند (مصـباح یزدی، ۱۳۹۴ الف، ص ۳۳۶). از طرفی، در این دیدگاه مقصود از قُرب، نه قُرب اعتباری و نه قُرب فلسفی است، بلکه منظور قُربی است که علم و ارادهٔ انسان نقش اساسی را در آن ایفا می‌کند، یعنـی قُربـی است که محصول و معلول آگاهی و شناخت و ارادهٔ انسان است (مصباح یزدی، ۱۳۹۴ الف، ص ۳۴۷).

در عصر حاضر، دستیارهای هوشمند و سیستم‌های خبره در جهت‌گیری افعال اختیـاری انسـان نقش مؤثری را ایفا می‌کنند. با توجه به تأثیر هوش مصنوعی بر آگاهی و معرفـت انسـان و بازتـابش در اراده او امروزه این ابزار بـه عنـوان عمـده‌تـرین ابـزار محـرک خـارجی و عامـل جـاذب بیرونـی محسوب می‌شوند که می‌تواند موجب کمالی نسبی برای نفس در مرتبه و شأنی خـاص از او شـود و می‌تواند در رسیدن نفس انسان به کمال حقیقـی معـد نفـس و معـین انسـان باشـد. بـرای مثـال، سیستم‌هایی مانند «تیک‌تاک» با تحلیل رفتار کاربر و متناسب با ذائقه او محتواهـایی را بـه صـورت هدفمند به او نمایش می‌دهند و ذهـن او را بـا خـود همـراه مـی‌کننـد. هدفمنـدی ایـن سیسـتم‌هـا می‌تواند مقُرب انسان به کمال حقیقی او باشـد و بـا ارائـه مسـیرهای تکـاملی شخصی‌سازی‌شـده متناسب با ویژگی‌های هر فرد، او را در شرایط مختلف در انتخاب بهترین مسیر هدایت کند.

در اخلاق با طراحی، تأکید بر آن است تا الگوریتم‌های هوش مصنوعی صرفاً بر پایهٔ کارآمـدی طراحی نشوند، بلکه با پیاده‌سازی سیستماتیک و جـامع ملاحظـات اخلاقـی در فراینـد طراحـی و توسعه سیستم‌های هوش مصنوعی زمینه‌های تکامل اخلاقی کاربران نیز فـراهم آیـد و در نتیجـه از مخاطرات احتمالی جلوگیری شود. بـه عنـوان مثـال، در دسـتیارهای صـوتی در تصـمیم‌گیـری در موقعیت‌های اضطراری، میزان حفظ حریم خصوصی کاربران تا چه حدی لازم است؟ تصور کنید یک کاربر به دستیار صوتی هوشمند خود بگوید: «من قصد خودکشی دارم». آیا هـوش مصـنوعی باید او را از این کار منصرف کند و یا بی‌تفاوت بماند و به حریم خصوصی کاربر احترام بگذارد؟ یـا به‌طور خودکار با اورژانس یا نزدیکان فرد تماس بگیرد؟ این یک معضل اخلاقی واقعی است کـه در آن هوش مصنوعی مجبور است بین حریم خصوصی و حفظ جان انسان‌ها یکی را انتخاب کند. از این‌رو، طراحی الگوریتم‌های اخلاقی در هوش مصنوعی باید بر مبنای اصـول و قواعـدی صـورت گیرد که استفاده از این فناوری در راستای تعالی انسان و تکامل همه‌جانبه او باشد.

بر این اساس و با توجه به مبانی اخلاقی علامه مصباح یزدی، می‌توان الگـوی مطلـوب نسـبت

اخلاق با هوش مصنوعی در سطح اخلاق با طراحی را با عنوان «غایت‌گرایی زمینه‌ساز کمال» ارائه کرد و بایسته است مهندسان و اخلاق‌پژوهان اسلامی در حوزهٔ اخلاق هنجاری هوش مصنوعی، در طراحی الگوریتم‌هایی که قابلیت‌های استدلال اخلاقی را به ابزارهای هوشمند اضافه می‌کنند، اصول و قواعدی را که در استفاده انسان از این ابزارها برای حصول کمال حقیقی معد و مُعین هستند، در نظر داشته باشند. بنابراین، با درک بهتر از اخلاق انسانی این نتیجه حاصل می‌شود که اخلاق ماشینی نتیجه شبیه‌سازی اخلاق انسانی نیست، بلکه نتیجهٔ یک اختراع اخلاقی یا به عبارتی، یک اخلاق مصنوعی است، اما این اخلاق می‌تواند حتی برای ارتقای اخلاق انسانی و کسب کمال نهایی به عنوان یک ابزار مصنوع، مناسب باشد.

بنابراین، در تعیین هدف و غایت برای این ابزارها می‌توان ناظر بر مکتب اخلاقی علامه مصباح یزدی در شیوهٔ تلفیقی، اصول بالا به پایینی را در هدایت رفتار عامل هوشمند توصیف کرد که ناظر بر هدف غایی انسان باشد و در رویکرد پایین به بالا ملزم به رجوع به وحی و تعیین مصادیق حُسن و قُبح و نمونه‌های کاربردی آن بود، اما باید دید پیاده‌سازی الگوریتم‌های اخلاقی متناسب با این نظریه، چه الزاماتی را می‌طلبد و مبنایی را باید اتخاذ کرد تا به این هدف دست یافت.

۸. الزامات رویکرد «غایت‌گرایی زمینه‌ساز کمال»

در این بخش به جهت وضوح نظریه «غایت‌گرایی زمینه‌ساز کمال» به توصیفات عینی‌تری از این رویکرد می‌پردازیم. برای رسیدن به این وضوح توضیح برخی اصطلاحات ضروری است.

۸. ۱. دو رهیافت عمده در هوش مصنوعی

در علوم شناختی ما با دو نگاه محاسبه‌گرایی و پیوندگرایی به ذهن مواجه هستیم که به موازات این دو نگاه، دو گونه هوش مصنوعی در نیمهٔ دوم قرن بیستم شکل گرفته است: رهیافت نمادگرایی و رهیافت اتصال‌گرایی.

۱. پارادایم نمادگرایی[1]

این رویکرد مبتنی بر نظریهٔ محاسبه‌گرایی است و ذهن را یک نظام قیاسی استنتاجی می‌بیند و تبیین ساختار

1. Symbolic Paradigm

زبان را در بازنمایی صریح دانش در قالب نمادها و قواعد صوری کافی می‌داند. سیستم‌های مبتنی بر منطق گزاره‌ای یا حساب محمولات، نمونه‌های بارز این رویکرد هستند. در این معماری، استنتاج از طریق اعمال قواعد صوری بر روی نمادها انجام می‌پذیرد. برای مثال، ماشین وضعیت-عمل[1] بر اساس مجموعه‌ای از قواعد «اگر/آنگاه» طراحی شده است که تعیین می‌کند در چه موقعیتی چه اقداماتی را انجام دهد و با تغییر موقعیت‌ها، چه به دلیل اقدامات قبلی و چه به دلیل تغییرات در محیط این قواعد اقدامات بعدی (یا عدم اقدام) را تعیین می‌کند (Russell & Norvig, 2003, p. 46)

۲. رهیافت اتصال‌گرایی[2]

این رویکرد که مبنای شبکه‌های عصبی عمیق است، مبتنی بر یادگیری استقرایی از طریق تنظیم پارامترهای وزنی در یک شبکه توزیع‌شده از واحدهای پردازشی است و سعی در شبیه‌سازی ساختار و کُنش مغز دارد. در این مُدل، دانش به صورت توزیع‌شده و ضمنی در الگوی اتصالات شبکه رمزگذاری می‌شود. برای نمونه، سیستم‌های تشخیص تصویر مبتنی بر یادگیری، از طریق پردازش هزاران نمونه، به صورت خودکار ویژگی‌های مرتبط را استخراج می‌کنند. در یادگیری ماشین، هدف اصلی تخصیص و تعمیم است؛ بدین معنا که نمونه‌های داده‌ای با توابعی که بر اساس این نمونه‌ها آموزش داده می‌شوند، به گروه‌های مختلفی اختصاص می‌یابند و سیستم روی نمونه‌های داده‌ای که در نظر گرفته نشده‌اند نیز به‌خوبی عمل می‌کنند (فرندنبرگ و سیلورمن، ۱۴۰۰، ص ۵۰۴). انواع یادگیری در چارچوب اتصال‌گرایی عبارتند از:

۱) یادگیری با نظارت:[3] در یادگیری با نظارت، یادگیری یک الگوی کلی ست که ورودی‌ها را به خروجی می‌برد. در این شیوه، از طریق پردازش هزاران نمونه برچسب‌دار، پاسخ مطلوب به سیستم آموزش داده می‌شود.

۲) یادگیری بدون نظارت:[4] این روش زمانی به کار می‌رود که داده‌های آموزشی فاقد برچسب باشند. سیستم باید به تنهایی ساختارهای نهفته در داده‌ها را کشف کند. هرچند اطلاعات خوشه‌بندی شده در این روش ممکن است جامع و کامل باشند، اما شناسایی الگوی موجود در این

1. Situation-Action Machine
2. Connectionist Paradigm
3. Supervised Learning
4. Unsupervised Learning

داده‌ها به راحتی امکان‌پذیر نیست. در یادگیری بانظارت و بدون نظارت، سیستم‌ها ممکن است سوگیری‌های موجود در داده‌های آموزشی و یا تعلیم یافته در تعامل با کاربران را بازتولید کنند.

۳) یادگیری تقویتی:[1] در این روش، عامل هوشمند از طریق تعامل با محیط و دریافت بازخورد به شکل پاداش یا تنبیه یاد می‌گیرد. در این حالت، الگوریتم‌های یادگیری باید قادر به انتخابِ سیاستِ مناسب باشد. سیستم‌های بدون مُدل و دارای مُدل در یادگیری تقویتی می‌توانند وجود داشته باشند. تمایز بین این دو می‌تواند یک روش محاسباتی بین اهداف و عادات در رفتارهای انسان ارائه کند (O'Doherty et al,, 2015, pp. 94-100).

۸. ۲. انتخاب رهیافت مناسب در «غایت‌گرایی زمینه‌ساز کمال»

بر اساس آنچه گفته شد، در پیاده‌سازی الگوریتم‌های اخلاقی نمی‌توان در تعیین مصادیق و نمونه‌های عملی به طور مطلق بر روش یادگیری متکی بود و از نظریه اتصال‌گرایی[2] که مبنای شبکه‌های عصبی مصنوعی است، استفاده کرد؛ زیرا مبنای استدلال در این روش یادگیری بر اساس آموزه‌های آموزشی و تجربه است.، بلکه مناسب‌تر این است که از نظریه نمادگرایی که بر اساس محاسبه‌گرایی و قوانین منطقی، دست به استنتاج اخلاقی می‌زند بهره برد و با توجه به مزیّت‌ها و چالش‌های انواع روش‌های یادگیری باناظر، یادگیری بدون ناظر و یا یادگیری تقویتی روشی را اتخاذ کرد که مبنای ارزش‌گذاری اخلاقی را در رویکرد پایین به بالا وحی قرار دهد.

مبتنی بر نظریه «غایت‌گرایی زمینه‌ساز کمال» به عنوان یک نمونه عینی، در طراحی سیستم تخصیص منابع مالی اسلامی می‌توان رویکردی جامع اتخاذ کرد که در سه سطح به هم‌پیوسته عمل کند. در سطح کلان (رویکرد بالا به پایین)، سیستم با تکیه بر اصول عقلی که نتیجه رویکرد محاسبه‌گرایی و استنتاج منطقی است، چارچوب ارزشی خود را شکل می‌دهد. به عنوان مثال، اصل عقلی «عدالت خوب است» معیار ثابتی را برای تعیین شاخص‌های تخصیص منابع مالی قرار می‌دهد. از جملهٔ این شاخص‌ها، تعیین مصادیق خمس و زکات و انفال و ضابطه تصرّف در آن‌ها، اولویت‌بندی مبتنی بر نیازهای اساسی، رفع تبعیض‌های طبقاتی و جنسیتی و توجه توأمان به مسئولیت‌های فردی و اجتماعی است که همگی ریشه در سیره نبوی و علوی دارند.

1. Reinforcement Learning
2. Connectionism

در سطح خُرد (رویکرد پایین به بالا) سیستم با بهره‌گیری از پیشرفته‌ترین تکنیک‌های پردازش زبان طبیعی و یادگیری عمیق، به تحلیل هزاران نمونه عینی از رفتار معصومین (ع) در مواجهه با مسائل اجتماعی و اقتصادی می‌پردازد. این تحلیل شامل بررسی مواردی مانند نحوه تقسیم امکانات توسط پیامبر اکرم (ص) در مدینه، رویه‌های قضایی امیرالمؤمنین (ع) در توزیع بیت‌المال و معیارهای تخصیص منابع مالی در کلام ائمه اطهار (ع) می‌شود. از این طریق، سیستم قادر است الگوهای رفتاری معصومین (ع) را در موقعیت‌های مختلف استخراج و مدل‌سازی کند.

چنانچه مکانیزم یادگیری سیستم به‌گونه‌ای طراحی شود که به صورت پویا و تحت نظارت کارشناسان دینی به روزرسانی شود، سیستم از سوگیری در نتیجه تعامل با کاربران غیراخلاقی مصون خواهد ماند. این مکانیزم ترکیبی از سه لایه اصلی است: موتور استنتاج منطقی که اصول کلی عقلی را مبنای استدلالات اخلاقی قرار می‌دهد، شبکه عصبی برمبنای یادگیری استقرایی که مصادیق جدید را با متون دینی مقایسه می‌نماید و سیستم یادگیری تقویتی که با دریافت بازخورد از کارشناسان دینی، مداوم خود را اصلاح و بهینه می‌کند.

حاصل این معماری پیشرفته، سیستمی هوشمند است که در عین بهره‌گیری از آخرین دستاوردهای فناوری، کاملاً در چارچوب ارزش‌های اسلامی و سیره معصومین (ع) عمل می‌کند. این سیستم نه‌تنها قادر است در توزیع منابع به عدالت رفتار کند، بلکه با الهام از سیره نبوی و علوی، در تشخیص اولویت‌ها و برخورد با موارد پیچیده نیز به عنوان الگویی عملی از «غایت‌گرایی زمینه‌ساز کمال» عمل می‌کند. بدین ترتیب، فناوری هوش مصنوعی در این سیستم به ابزاری برای تحقّق عدالت اجتماعی و تُقرب‌الهی تبدیل می‌شود.

۹. نتیجه‌گیری

نگرانی‌های اخلاقی بوجود آمده از رشد سریع سیستم‌های متکی بر هوش مصنوعی زمینه‌ساز شکل‌گیری دانش میان‌رشته‌ای «اخلاق هوش مصنوعی» گردید تا در دو بخش طراحی سیستم‌های هوشمند و تعامل انسان با ماشین در جهت بهبود وضعیت و کاستن از دغدغه‌های اخلاقی به نقد و نظر بپردازد.

با توجه به گسترهٔ این دانش و سطح‌بندی‌های مختلفی که در نسبت اخلاق با هوش مصنوعی صورت گرفته است، رویکرد این پژوهش تنها بر سطح «اخلاق با طراحی» (یا اخلاق آمیخته با

طراحی) معطوف است که در آن به بخش طراحی سیستم‌های هوشمند و ادغام فنی الگوریتمی قابلیت‌های استدلال اخلاقی به عنوان بخشی از رفتار سیستم‌های مستقل پرداخته می‌شود و بنابراین، در صدد تبیینی از صورت‌بندی الگوی مطلوب در نسبت اخلاق با طراحی از دیدگاه علامه مصباح یزدی (ره) برآمده است. بدین منظر با طرح سه نظریه رایج «اصل‌گرایی»، «نمونه‌گرایی» و «تعادل‌تأملی» که در فرایند معرفت اخلاقی در انسان مطرح است؛ به سه رویکرد «بالا به پایین»، «پایین به بالا» و «تلفیقی» در پیاده‌سازی اخلاق در ماشین پرداخته شد.

از این‌رو، با التفات به نظر علامه مصباح یزدی (ره) در نهایت این جمع‌بندی حاصل شد که می‌توان دیدگاه اخلاقی ایشان را با تبیینی ویژه و تأکید بر تمایزات موجود، ذیل نظریه «تعادل تأملی» مندرج دانست و در مرحلهٔ کاربست اخلاق در ماشین نیز آن را ذیل رویکرد «تلفیقی» به عنوان الگوی مطلوب در میان رویکردهای موجود در نظر گرفت و از اخلاق با طراحی برای ارتقای اخلاق انسانی و کسب کمال نهایی به عنوان یک ابزار مصنوع، استفاده کرد. چگونگی پیاده‌سازی این رویکرد در سیستم‌های هوشمند پژوهشی مستقل را می‌طلبد که در آن به همه ابعاد فنی و مبنایی آن پرداخته شود.

فهرست منابع

جمشیدیها، غلامرضا؛ عبادی، زینب. (۱۳۹۹). بررسی عقلانیت ارزش‌شناختی و چگونگی حصول آن در اندیشه ماکس وبر «تجزیه و تحلیل منطقه خاکستری»، مجله جامعه‌شناسی ایران. ش ۳، ص۲۴ـ۵۸.

ربّانی گلپایگانی، علی. (۱۳۸۳). مدرنیته و عقلانیت. مجله تخصصی کلام اسلامی، ش۴۸، ص۳۹ ـ ۵۸.

عبدخدایی، زهره و توحیدی، نسیم و کریمی واقف، نرگس و براتی، محمد. (۱۴۰۰). هوش سفید؛ از اخلاق ماشینی تا ماشین اخلاقی. تهران: مؤسسه مطالعات فرهنگی و اجتماعی.

فرندنبرگ جی؛ سیلورمن، گوردن. (۱۴۰۰)، علوم شناختی: مقدمه‌ای بر مطالعه ذهن. (ترجمه: محسن افتاده‌حال و دیگران، چاپ سوم). تهران: مؤسسه آموزشی پژوهشی صنایع دفاعی.

مصباح یزدی، محمد تقی. (۱۳۹۴الف). نقد و بررسی مکاتب اخلاقی. قم: انتشارات مؤسسه آموزشی و پژوهشی امام خمینی (ره).

مصباح یزدی، محمدتقی. (۱۳۹۴ب). فلسفهٔ اخلاق. (تحقیق و نگارش: احمد حسین شریفی). قم: انتشارات مؤسسه آموزشی و پژوهشی امام خمینی (ره).

Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, *7*(3), 149–155. https://doi.org/10.1007/s10676-006-0004-4

Aristotle. (n.d.). *Nicomachean ethics* (W. D. Ross, Trans.). MIT Classics. Available at: https: //classics.mit.edu/Aristotle/nicomachaen.2.ii.html.

Brey, P., & Dainow, B. (2023). Ethics by design for artificial intelligence. AI and Ethics. Advance online publication. https://doi.org/10.1007/s43681-023-00345-3

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, 20(1), 1–3.

Giarmoleo, F. V., Ferrero, I., Rocchi, M., & Pellegrini, M. M. (2024). What ethics can say on artificial intelligence: Insights from a systematic literature review. *Business and Society Review*. Advance online publication.

Govindarajulu, N. S., & Bringsjord, S. (2017). On Automating the Doctrine of Double Effect. *Rensselaer Polytechnic Institute*. Troy, NY, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Pages 4722-4730.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines.*Springer: Minds and Machines. Journal for Artificial Intelligence*. Philosophy and Cognitive Scienc. Volume 30, pages 99–120.

Leikas, J., Koivisto, R., & Gotcheva, N. (2019). Ethical framework for designing autonomous intelligent systems. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(1), 18.

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2012). Robot ethics: The ethical and social implications of robotics. MIT Press.

O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). T*he structure of reinforcement-learning mechanisms in the human brain*. Current Opinion in Behavioral Sciences, 1, 94–100.

Osasona, F., Ogunbiyi, O., & Adebayo, A. (2024). Reviewing the ethical implications of AI in decision making processes. *International Journal of Management & Entrepreneurship Research*, 6(2), 322–335.

Picard, R. W. (1995). *Affective Computing*. MIT Press. Cambridge. MA.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Prentice Hall.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). *Implementations in machine ethics: A survey*. ACM Computing Surveys, 53(6), Article 132.

Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press.