فصلنامه علمی- پژوهشی اخلاق پژوهی
سال هشتم • شماره سوم • پاییز ۱۴۰۴
Quarterly Journal of Moral Studies
Vol. 8, No. 3, Autumn 2025

سنجش میزان رشد اخلاقی توسط دستیارهای محاسبه مبتنی بر هوش مصنوعی؛ چالش‌ها و راهبردها

مجید غلامی* | بهروز مینایی بیدگلی** | هادی حسین‌خانی***

10.22034/ethics.2025.52139.1814

چکیده

ظهور مدل‌های زبانی بزرگ (LLMs)، افق‌های تازه‌ای پیش روی دستیارهای هوشمند «محاسبه نفس» گشوده؛ با این حال، ماهیت انتزاعی مفاهیم اخلاقی، سنجش ماشینی آنها را با دشواری‌هایی مواجه ساخته است. این پژوهش، ضمن استخراج، تحلیل و طبقه‌بندی چالش‌های فنی و مفهومیِ سنجش اخلاق در بستر «دستیارهای هوشمند محاسبه نفس»، در صدد تدوین راهبردهای نظری برای مدیریت آنها است. یافته‌ها در چارچوب تحلیلی چهاربخشی تبیین شده‌اند: ۱) چالش‌های ترجمهٔ مفهومی به ساختار محاسباتی از قبیل عملیاتی‌سازی نیّت و فقدان هستان‌نگارهای اخلاقی؛ ۲) چالش‌های داده‌محور همچون اتکا به ردپاهای دیجیتال، کمبود مجموعه داده‌های استاندارد و سوگیری فرهنگی؛ ۳) چالش‌های منطق الگوریتمی نظیر ناپایداری مدل، مسئلهٔ جعبه‌سیاه و حساسیت به بیان؛ ۴) چالش‌های تعاملی و پویایی که شامل حلقه‌های بازخورد معیوب و خطای نسبت‌دهی زمانی است. پژوهش نتیجه می‌گیرد که غلبه بر این موانع مستلزم فراتر رفتن از رویکردهای صرفاً آماری است. راهبردها شامل توسعه نمایه‌های سنجش چندبعدی، طراحی معماری‌های ترکیبی با ادغام هوش مصنوعی تبیین‌پذیر، استنتاج علّی و هستان‌نگارهای رسمی مبتنی بر اخلاق اسلامی، و گذار از مشاهده منفعلانه به خودارزیابی فعالانه و کاربرمحور است.

کلیدواژه‌ها

محاسبه نفس، سنجش اخلاق، مدل‌های زبانی بزرگ، هوش مصنوعی، دستیارهای هوشمند محاسبه.

* دانشجوی دکتری مدرسی معارف اسلامی، مؤسسه آموزشی و پژوهشی امام خمینی (ره)، قم. ایران. (نویسنده مسئول) | yabager248@gmail.com
** استاد دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت تهران، تهران. ایران. | b_minaei@iust.ac.ir
*** استادیار گروه اخلاق مؤسسه آموزشی و پژوهشی امام خمینی (ره)، قم. ایران. | hkh@iki.ac.ir

تاریخ دریافت: ۱۴۰۴/۰۵/۲۶ ▣ تاریخ تأیید: ۱۴۰۴/۰۹/۰۵

■ غلامی، مجید؛ مینایی بیدگلی، بهروز؛ حسین‌خانی، هادی. (۱۴۰۴). سنجش میزان رشد اخلاقی توسط دستیارهای محاسبه مبتنی بر هوش مصنوعی؛ چالش‌ها و راهبردها. فصلنامه اخلاق‌پژوهی. ۸(۲۸)، ۵ ـ ۳۸ . doi: 10.22034/ethics.2025.52139.1814

۱. مقدمه

هوش مصنوعی و به‌ویژه مدل‌های زبانی بزرگ، به‌سرعت در حال تبدیل شدن به ابزارهای جدایی‌ناپذیر زندگی روزمرّه هستند. یکی از کاربردهای نوظهور و در عین حال چالش‌برانگیز این فناوری، استفاده از آن به‌عنوان «دستیار هوشمند محاسبه نفس» است؛ ابزاری که می‌تواند به افراد در مسیر خودسازی و رشد اخلاقی یاری رساند. با این حال، سنجش مفاهیم انتزاعی و الهی مانند نیت، اخلاص و فضایل اخلاقی، با موانع فنی و مفهومی جدّی روبه روست. چگونه می‌توان پدیده‌ای کیفی و درونی را به معیارهای کمّی و قابل محاسبه برای یک ماشین ترجمه کرد؟ این مقاله با هدف پاسخ به این پرسش، به شناسایی و تحلیل نظام‌مند چالش‌های مهم در این حوزه می‌پردازد. پژوهش حاضر، ابتدا موانع را در چهار دسته اصلی (ترجمه مفهومی، داده‌محور، منطق الگوریتمی و تعاملی) طبقه‌بندی کرده و سپس راهبردهای نظری جامعی را برای طراحی نسل آینده این دستیارها ارائه می‌دهد؛ راهبردهایی که بر گذار از رویکردهای صرفاً آماری به معماری‌های ترکیبی، تبیین‌پذیر و مبتنی بر دانش استوارند.

۲. پیشینه پژوهش

مرور پیشینهٔ پژوهش نشان می‌دهد که چالش‌های فنی و مفهومی سنجش اخلاق توسط هوش مصنوعی به صورت گسترده، اما پراکنده در حوزه‌های تخصصی گوناگون مورد بررسی قرار گرفته‌اند. پژوهشگران به موانع بنیادین در «عملیاتی‌سازی» مفاهیم انتزاعی اخلاق و ترجمه آن به ساختارهای محاسباتی اذعان کرده‌اند؛ شکافی که میان مفاهیم اخلاقی و نیاز سیستم‌های محاسباتی به ورودی‌های گسسته وجود دارد (Jacobs & Wallach, 2021, p. 375). در سطح داده‌ها، موانعی همچون کمبود مجموعه داده‌های استاندارد برای آموزش و ارزیابی مدل‌ها (Ji et al., 2025, p. 63) و سوگیری‌های فرهنگی و دینی در داده‌های آموزشی (Abrar et al., 2025, p. 1) مورد تأکید قرار گرفته‌اند. از منظر منطق الگوریتمی نیز ناپایداری ذاتی قضاوت مدل‌ها (Atil et al., 2025, p. 1) و خطرات ناشی از حلقه‌های بازخورد معیوب که می‌توانند سوگیری‌های اولیه را تشدید کنند، شناسایی شده‌اند (Krauth et al., 2022, p. 2). این حلقه‌های بازخورد در تعامل انسان و هوش مصنوعی بسیار قوی‌تر عمل می‌کنند و قادرند خطاهای جزئی را به سوگیری‌های کلان تبدیل کنند (Glickman & Sharot, 2024, pp. 345-346).

با این وجود، خلأ اصلی، فقدان یک چارچوب تحلیلی یکپارچه و نظام‌مند است که این چالش‌های متنوع را که در حوزه‌های تخصصی مجزّا (مانند پردازش زبان طبیعی، یادگیری ماشین و هوش مصنوعی تبیین‌پذیر) بررسی شده‌اند، در یک سِیر منطقی (از تعریف مفهومی تا تعامل پویا) طبقه‌بندی کرده و به‌طور خاص بر مسئله «سنجش اخلاق» متمرکز شود. علاوه بر این، کاربرد این چالش‌ها در زمینهٔ تخصصی «دستیارهای هوشمند محاسبه نفس» مبتنی بر چارچوب اخلاق اسلامی، که در آن مفاهیمی چون «نیت» و «ترک فعل» نقشی محوری دارند، حوزه‌ای بِکر و بررسی‌نشده باقی مانده است. پژوهش حاضر با هدف پر کردن این خلأ، تلاش می‌کند تا با ارائه یک چارچوب تحلیلی چهاربخشی، به شناسایی، طبقه‌بندی و تحلیل نظام‌مند این چالش‌ها پرداخته و راهبردهای جامعی برای غلبه بر آنها، با الهام از مبانی اخلاق اسلامی و روش‌های نوین هوش مصنوعی، ارائه کند.

۳. روش پژوهش

این پژوهش با رویکرد تحقیق اِسنادی و با ماهیتی کیفی و توصیفی-تحلیلی، به شناسایی، طبقه‌بندی و تحلیل چالش‌های فنی و مفهومی سنجش اخلاق توسط دستیارهای هوشمند و ارائه راهبردهای نظری برای آنها می‌پردازد. فرایند پژوهش در چند مرحله نظام‌مند اجرا شد. اگر چه این مراحل به صورت متوالی ارائه شده‌اند، فرایند اجرا در عمل به صورت چرخه‌ای و بازتابی[1] بود: نتایج هر مرحله، به‌ویژه تحلیل مضمون، به‌طور مکرر مبنای بازنگری در مراحل پیشین، اصلاح پرامپت‌ها و پالایش معیارهای جست‌وجو قرار گرفت تا به اشباع نظری و پایداری مضامین دست یابیم.

۳. ۱. مرحلهٔ اول: شناسایی و پالایش اولیه چالش‌ها

در مرحلهٔ اکتشافی، با استفاده از مدل‌های زبانی بزرگ و بارش فکری، دامنه وسیعی از چالش‌های بالقوه استخراج شد. سپس، این چالش‌ها بر اساس معیارهای دقیقی پالایش شدند: الف) چالش‌هایی که عمومی بوده و بین سنجش توسط انسان و هوش مصنوعی مشترک هستند،

1. iterative and reflexive

حذف شدند؛ ب) چالش‌هایی که به طور کلی به همه انواع دستیارهای هوشمند مربوط می‌شوند و مختص دستیارهای سنجش اخلاق نیستند، کنار گذاشته شدند؛ ج) برای تفکیک میان دو حوزه پژوهشی متمایز، مقالات متمرکز بر «اخلاقیات هوش مصنوعی» (طراحی دستیارهایی که به اصول اخلاقی پایبند هستند) از مقالات مرتبط با «سنجش اخلاق انسان توسط هوش مصنوعی» جدا شدند تا تمرکز پژوهش حفظ شود؛ د) دغدغه‌هایی که از تبعات استفاده از دستیارها هستند، نظیر پیامدهای روان‌شناختی (مانند وابستگی)، مسائل کلان حقوقی (مانند حریم خصوصی و شفافیت) یا مباحث مرتبط با سیاست‌گذاری و تعیین وظایف نهادها، کنار گذاشته شدند تا تمرکز بر چالش‌های فنی و مفهومیِ خودِ فرایند سنجش باقی بماند. این فرایند به تحدید دقیق دامنه مسئله و تمرکز بر چالش‌های تخصصی این حوزه کمک کرد.

۳. ۲. مرحلهٔ دوم: گردآوری و نمونه‌گیری هدفمند منابع

گردآوری ادبیات پژوهشی، با گفت‌وگوهای اکتشافی با مدل‌های زبانی بزرگ (LLMs) برای شناسایی منابع کلیدی آغاز شد و در ادامه، با کاوش در پایگاه‌های داده معتبر علمی نظیر arXiv.org، ACM Digital Library، و Google Scholar ادامه یافت. این فرایند با استفاده از پرامپت‌های مهندسی‌شده و روش نمونه‌گیری انباشتی[1] برای گسترش نظام‌مند منابع انجام شد. گزینش منابع بر اساس معیارهایی چون به‌روز بودن (با اولویت سال‌های ۲۰۲۳ تا ۲۰۲۵ میلادی)، ارتباط و نوآوری صورت گرفت تا به اشباع نظری نسبی دست‌یابیم.

۳. ۳. مرحلهٔ سوم: تحلیل، سنتز و طبقه‌بندی

پس از گردآوری منابع، با بهره‌گیری از روش «تحلیل مضمون»[2] چالش‌ها و راهبردها از متون استخراج شدند. این فرایند با «تولید کدهای اولیه» آغاز شد، مطابق راهنمای شش‌مرحله‌ای براون و کلارک (Braun & Clarke, 2006, p. 87). کدها سپس پالایش و ادغام شدند تا مضامین بالقوه پدید آیند. رویکرد اتخاذ شده در این تحلیل، یک تحلیل مضمون انعطاف‌پذیر[3] ست. بنابراین،

1. Snowball Sampling
2. Thematic Analysis
3. reflexive thematic analysis

چارچوب نهـایی چهاربخشـی ایـن مقالـه (ترجمـه مفهـومی، داده‌محـور، منطـق الگـوریتمی و تعاملی) محصول سنتز تحلیلی است که چالش‌ها را بر اساس سِیر منطقیِ تبـدیل یـک مفهـوم انتزاعی به یک سیستم تعاملی سازمان‌دهی می‌کند؛ این سِیر از تعریـف مفهـومی آغـاز شـده، بـه پردازش داده‌ها و منطق مُدل می‌رسد و در پایان به تعامل با کاربر ختم می‌شود.

۴. چارچوب نظری و مفهوم‌شناسی

در ابتدا، مفاهیم اصلی این مقاله بـا الهـام از چارچوب‌هـای اخـلاق اسـلامی و تطبیـق آنهـا بـا فناوری‌های نوین تعریف می‌شوند.

دستیار هوشمند محاسبه نفس: برای تعریف این مفهوم، باید سه جزء اصلی آن را تعریف کرد:

هوش مصنوعی (AI): این اصطلاح، دارای تعاریف متعددی است که دو رویکـرد اصـلی در میان آنها برجسته‌تر است: تمرکز بـر «عملکـرد شبه‌انسـانی» کـه ریشـه در آزمـون تورینـگ دارد (Turing, 1950, p. 433) و تمرکز بر «عقلانیت». رویکرد اول بر تقلید رفتار انسـان تأکیـد دارد، امـا رویکرد دوم که به «مدل استاندارد» در هـوش مصـنوعی شـهرت یافتـه، ایـن حـوزه را مطالعـه و ساخت عامل‌های هوشمند[1] تعریـف می‌کنـد (Russell & Norvig, 2021, p. 22). در ایـن پـژوهش، تعریف دوم، یعنی هوش مصنوعی به مثابه ساخت عامل‌های عقلانی، به عنوان زیربنای مفهـومی برگزیده شده است. دلیل این انتخاب آن است که هدف «دستیار هوشمند محاسبه نفـس» صـرفاً تقلید مکالمه نیست، بلکه طراحی سیستمی است که محـیط خـود (داده‌هـای رفتـاری کـاربر) را درک کرده و برای دستیابی به یک هدف مشخص (تسهیل تأمل و خودآگاهی اخلاقی)، اقـداماتی بهینه را انجام می‌دهد. بر این اساس، عامل هوشمند، سیستمی است که با اسـتفاده از حس‌گرهـا به درک محیط پرداخته و از طریق عملگرها، برای رسیدن به بهترین نتیجـه ممکـن (یـا در شـرایط عدم قطعیت، بهترین نتیجه مورد انتظار) اقدام می‌کند. این دیـدگاه، چـارچوب نظـری دقیـق‌تری برای تحلیل چالش‌های طراحـی دسـتیار هدفمنـد و مفیـد فـراهم مـی‌آورد و قابلیت‌هـایی نظیـر پردازش زبـان طبیعـی، بازنمـایی دانـش، اسـتدلال خودکـار و یـادگیری ماشـینی را در خـدمت بهینه‌سازی یک تابع هدف مشخص (رشد اخلاقی کاربر) قرار می‌دهد.

1. Intelligent Agents

دستیارهای هوشمند:[1] این دستیارها سیستم‌هایی اطلاعاتی هستند که برای کمک بـه کـاربران در انجام وظایف و اهدافشان طراحی شده‌اند (Dhiman et al., 2022, p. 645). چنـین دسـتیارهایی را می‌توان نوعی «عامل هوشمند» دانست که به مثابه همکار شناختی کاربر عمـل می‌کننـد و دارای ویژگی‌هایی مانند واکنش‌پـذیری، خودمختـاری و هـدف‌گرایی هسـتند (Wooldridge & Jennings, 1995, p. 116). نسل جدید این دستیارها که مبتنی بر مدل‌های زبـانی بـزرگ (LLMs) هسـتند، در قالب «Personal LLM Agents» تعریف شده‌اند، یعنی عامل‌هایی که به‌طور عمیـق بـا داده‌هـا، دستگاه‌ها و سرویس‌های شخصی کـاربر یکپارچـه شـده و از طریـق زبـان طبیعـی، کمک‌هـای هوشمندانه و شخصی‌سازی‌شده برای کاهش کارهای تکراری و تمرکز کاربر بر امور مهم‌تر فراهم می‌کنند (Li et al., 2024, p. 1).

محاسبه: شالوده نظری این پـژوهش بـر تلفیـق فراینـد چهارمرحلـه‌ای «مرابطـه» در اخـلاق اسلامی، یعنی مشارطه، مراقبه، محاسبه و معاقبه (مصباح یزدی، ۱۳۹۲، ص ۳۵) با قابلیت‌هـای هـوش مصـنوعی اسـتوار اسـت. دسـتیار هوشـمند می‌توانـد نقـش ابـزاری فناورانـه بـرای تقویـت و دیجیتالی‌سازی هر یک از این مراحل را ایفا کند.

مشارطه: در این مرحله، فرد با خود عهد می‌بندد که از سرمایه عمر برای اهداف اخلاقـی بهـره گیرد (امام خمینی، ۱۳۸۰، ص ۹). نقش دستیار، تبدیل این «قصد ذهنی» به «برنامه عملیاتی» اسـت و با قابلیت‌های برنامه‌ریزی و یادآوری، به کاربر در تعریف و پیگیری اهداف اخلاقی (مانند «امروز خشم خود را کنترل کنم») کمک می‌کند.

مراقبه: این مرحله به معنای نظارت دائمی بر عهد اولیه در طول روز اسـت (مصـباح یـزدی، ۱۳۹۶، ج ۱، ص ۳۶۹). نقش دستیار، ایفای نقش یک «همراه مراقـب» دیجیتـال اسـت کـه بـا اعلان‌هـای هوشمند برای خوداظهاری یا تحلیل داده‌های رفتاری، به حفظ آگاهی کاربر نسـبت بـه اهـدافش کمک می‌کند.

محاسبه: این مرحله به «حسابرسی» اعمال در پایان دوره می‌پردازد (غزالـی، بی‌تـا، ج ۱۵، ص ۲۵). در این مقاله، «سنجش» معادل فنی و عملیاتی «محاسـبه نفـس» اسـت. نقـش دسـتیار، تبـدیل «تأمل کیفی» به «تحلیل داده‌محور» اسـت. سیسـتم بـا پـردازش داده‌هـا، گزارش‌هـای بصـری، الگوهای رفتاری و یک «کارنامه اخلاقی» ارائه می‌دهد. سنجش دقیق در تمام مراحل مرابطـه، از اولویت‌بندی تعهدات تا تناسب‌سنجی بازخورد، نقشی محوری دارد.

1. Intelligent/Smart Assistants

معاقبه: بر اساس نتایج محاسبه، فرد نفس خود را سرزنش کرده یا تنبیهی برای آن در نظر می‌گیرد (مکارم شیرازی و همکاران، ۱۳۸۵، ج ۱، ص ۲۶۱). نقش دستیار در اینجا ارائه «بازخورد اصلاحی» و «تقویت مثبت» است. سیستم می‌تواند راهکارهایی برای جبران قصورها پیشنهاد دهد یا با برجسته‌سازی موفقیت‌ها، رفتار مطلوب را تقویت کند. نتایج این مرحله، ورودی مرحله مشارطه برای دوره بعدی خواهد بود و حلقه بازخورد را تکمیل می‌کند.

بر این اساس، «دستیار هوشمند محاسبه نفس» نوعی عامل هوشمند شخصی و تخصصی است که مبتنی بر چارچوب‌های اخلاق اسلامی، برای یاری رساندن به کاربر در فرایند سنجش و تأمل در صفات، رفتارها و نیات اخلاقی طراحی شده است. هدف غایی این سیستم، صدور قضاوت قطعی نیست، بلکه تسهیل محاسبه نفس و رشد اخلاقی از طریق بازخوردهای هوشمندانه و پرسش‌های تأمل‌برانگیز است.

بنابراین، «دستیار هوشمند محاسبه» وظیفه تسهیل و دیجیتالی‌سازی مرحله «محاسبه» را به عنوان یکی از ارکان چهارگانه مرابطه بر عهده دارد. با این وجود، چنین دستیاری دربردارنده یک معماری تعاملی است که کل فرایند مرابطه را به یک سیستم سایبرنتیک برای خودسازی تبدیل می‌کند. تمرکز اصلی چالش‌های فنی این مقاله بر مرحله «محاسبه» قرار دارد؛ زیرا بر اساس الگوی چهارمرحله‌ای، این مرحله خروجی مراحل قبل را ارزیابی کرده و مبنای مرحله بعد (معاتبه) قرار می‌گیرد (مصباح یزدی، ۱۳۹۲، ص ۳۵). در نتیجه، در سیستمی هوشمند، موفقیت کل چرخه به دقت و اعتبارسنجی در این مرحله وابسته است.

۵. یافته‌ها: چارچوب تحلیلی چالش‌ها

یافته‌های پژوهش در قالب یک چارچوب تحلیلی چهاربخشی ارائه می‌شود که چالش‌های تخصصی سنجش اخلاق توسط دستیارهای هوشمند را طبقه‌بندی و تشریح می‌کند.

۵. ۱. چالش‌های ترجمه مفهومی به ساختار محاسباتی

این دسته از چالش‌ها به مشکلات بنیادین در «عملیاتی‌سازی» مفاهیم انتزاعی اخلاق می‌پردازند. این ترجمه، شکاف میان غنای مفاهیم اخلاقی و نیاز سیستم‌های محاسباتی به ورودی‌های گسسته را آشکار می‌سازد. هر گونه خطا در این سطح، به لایه‌های بعدی منتقل شده

و اعتبار کل سیستم را تضعیف می‌کند.

۵. ۱. ۱. نبود هستان‌نگارها برای سنجش اخلاق

یکی از بنیادی‌ترین موانع، فقدان هستان‌نگارها[1] یا به عبارت ساده‌تر چارچوب‌های مفهومیِ صوری و قابل پردازش از مفاهیم اخلاقی است. بدون چنین چارچوبی، مدل‌ها قادر به درک معنای دقیق فضایل و رذائل نیستند. بیشتر مدل‌های هوش مصنوعی تنها از هم‌رخدادی زبانی برای استنتاج مفاهیم اخلاقی بهره می‌برند، یعنی یاد می‌گیرند «چگونه مردم دربارهٔ اخلاق سخن می‌گویند»، نه آنکه «اخلاق چه ساختاری دارد» (Aijaz et al., 2025, p. 1). این رویکرد آماری سبب می‌شود ارزیابی اخلاقی سیستم‌ها فاقد بنیان هنجاری باشد. برای مثال، یک دستیار بدون هستان‌نگار، «بخشش» را صرفاً با کلیدواژه‌ها می‌شناسد، اما قادر به تمایز میان «بخشش از سر اخلاص» و «بخشش برای ریا» نیست؛ زیرا این تمایز نیازمند درک نیّت و زمینه است که تنها یک هستان‌نگار فراهم می‌کند.

۵. ۱. ۲. چالش تبدیل داده به سنجش و قضاوت

تبدیل ردِّ پاهای دیجیتال کاربر به یک سنجش اخلاقی معتبر، چالشی دیگر است که مستلزم انتخاب‌های انسانی و هنجاری است؛ از جمله وزن‌دهی به ابعاد مختلف رفتار و نحوه مدیریت عدم‌قطعیت مُدل در قضاوت‌ها. افزون بر این، تلاش برای فروکاستن چنین فرایند پیچیده‌ای به یک امتیاز عددی واحد، ساده‌انگارانه است (Nie et al., 2023, p. 1). این پیچیدگی در نظام اخلاق اسلامی ابعاد عمیق‌تری دارد. در این دیدگاه، درجه ارزش یا ضد ارزش بودن یک عمل به‌سادگی قابل اندازه‌گیری نیست و محصول تعامل پیچیده مجموعه‌ای از معیارهاست (عالم‌زاده نوری، ۱۳۹۶، ص ۳۲۹-۳۶۰). این معیارها را می‌توان به دو دسته کلی تقسیم کرد:

۱. معیارهای با جنبهٔ کمّی برجسته‌تر: معیارهایی نظیر «مقدار عمل» (تعداد، زمان، حجم)، «استمرار» (تداوم زمانی)، «زمان و مکان» خاص وقوع عمل، و «حُسن و قُبح فعلی» (ارزش ذاتی عمل) که قابلیت اندازه‌گیری عددی بیشتری دارند. هرچند این معیارها نیز ریشه‌های کیفی دارند (مثلاً قداست مکان)، اما سنجش آنها به شاخص‌های عینی نزدیک‌تر است.

1. Ontologies

۲. معیارهای با جنبه کیفی برجسته‌تر: معیارهایی نظیر «نیّت» (اخلاص)، «علم و آگاهی فاعل»، «میزان دشواری عمل» (که وابسته به شرایط فاعل است)، «اهمیت نسبی در تزاحم»، «گستره تأثیر اجتماعی»، «جامعیت عمل در سبک زندگی» و «زمینه‌سازی برای اعمال دیگر». این معیارها ذاتاً توصیفی و وابسته به زمینه هستند و تبدیل آنها به شاخص‌های عددی بسیار دشوار است. چالش اصلی این است که این معیارها مستقل از هم عمل نمی‌کنند و بر یکدیگر تأثیر متقابل دارند. بنابراین، محاسبه ارزش اخلاقی بیشتر شبیه به «قضاوت اخلاقی» است تا «محاسبه ریاضی»؛ زیرا نیازمند تحلیل روابط پیچیده میان این ابعاد در یک بستر خاص است. الگوریتم قضاوت باید منطقی برای شناسایی، وزن‌دهی و ترکیب این ابعاد داشته باشد تا به ارزیابی نهایی برسد، که این امر هوش مصنوعی را به یک قاضی هنجاری تبدیل می‌کند.

۵. ۱. ۳. تعریف فنی «واحد عملِ» قابل محاسبه

هر سیستم محاسباتی نیازمند تجزیه پدیده‌های پیوسته به واحدهای گسسته است. در سنجش اخلاق، این واحد بنیادین، «عمل»[1] است. تصمیم‌گیری دربارهٔ اینکه چه چیزی یک واحد عمل مستقل است (یک کلیک؟ یک پیام کوتاه؟) یک تقلیل‌گرایی فنی[2] است که ماهیت زمینه‌مند رفتار انسان را نادیده می‌گیرد. «عمل» تنها زمانی دارای معنای اخلاقی قابل تحلیل است که در بستر یک رویداد و همراه با اطلاعات زمینه‌ای تفسیر شود. تبیین عمل بدون در نظر گرفتن نقش عامل، نیّت اخلاقی، پیامد، و ویژگی‌های پیامد نظیر شدت، فایده و مدت‌زمان آن ممکن نیست (Aijaz et al., 2025, pp. 7-10). برای مثال، یک پیام متنی مانند «موافقم» به تنهایی بی‌اهمیت است، اما در پاسخ به پیشنهاد غیبت، به مصداق «اعانت بر اثم» تبدیل شده و بار اخلاقی منفی پیدا می‌کند. اتکای سیستم به واحدهای عملِ مُجزّا، آن را از درک این پویایی‌های زمینه‌ای بازمی‌دارد و مبنای محاسبات سنجش را بر پایه‌ای شکننده قرار می‌دهد.

۵. ۲. چالش‌های داده‌محور

این بخش بر موانع ناشی از خودِ داده‌ها تمرکز می‌کند؛ از نحوه بازنمایی واقعیت در داده‌های

1. Action
2. Technical Reductionism

دیجیتال تا سوگیری‌های موجود در فرایند جمع‌آوری و برچسب‌گذاری آنها. چالش‌های این حوزه نشان می‌دهند که حتی با وجود یک چارچوب مفهومی دقیق، کیفیت و محدودیت‌های داده‌های ورودی می‌تواند اعتبار سنجش را به‌طور جدّی تضعیف کند.

۵. ۲. ۱. وابستگی به شواهد قابل رصد دیجیتال

اساسی‌ترین محدودیت، ماهیت ناقص داده‌های دیجیتال است. ردِّ پاهای دیجیتال صرفاً «تصویری ناقص» از شخصیت، ارزش‌ها و رفتار افراد ارائه می‌کنند و بسیاری از ابعاد تجربهٔ انسانی ثبت نمی‌شود (Armstrong et al., 2023, p. 2). بر این اساس، در اخلاق اسلامی که «نیت» نقش محوری دارد، صرف ثبت عمل ظاهری (مثلاً کمک مالی آنلاین) برای سنجش ارزش اخلاقی (اخلاص یا ریا) کافی نیست؛ در نتیجه، هوش مصنوعی تنها بخشی از واقعیت را مشاهده می‌کند که اعتبار هر گونه سنجش اخلاقی را با چالش مواجه می‌سازد.

۵. ۲. ۲. اتکای مفرط بر پراکسی‌های سکویی

این چالش با اتکای مفرط بر «پراکسی‌ها»[1] یا نماگرهای ساده سکویی تشدید می‌شود. هوش مصنوعی برای سنجش مفاهیم پیچیده به سیگنال‌های دیجیتالی مانند «لایک» متکی است؛ در حالی که برطبق برخی پژوهش‌ها، لایک‌ها تنها بازتابی سطحی از حال‌وهوای لحظه‌ای کاربران‌اند (Alsabah, 2025, pp. 284–286)؛ برای مثال، آیا «لایک کردنِ» یک پست مفید، نماینده دقیقی برای فضیلت «امر به معروف» است، یا صرفاً همراهی با یک موج اجتماعی بدون نیّت اصلاحی است؟ در واقع، هوش مصنوعی تصویر ناقصی را که از واقعیت مشاهده می‌کند، از طریق لنزی ساده‌انگارانه تفسیر می‌کند.

۵. ۲. ۳. کمبود مجموعه داده‌های استاندارد اخلاقی

کمبود مجموعه داده‌های[2] استاندارد و معتبر، پیشرفت در حوزه سنجش اخلاق را کند کرده است. بدون چنین منابعی، آموزش، ارزیابی و مقایسه عادلانه مدل‌ها تقریباً غیر ممکن است.

1. Platform Proxies
2. Datasets

پژوهش‌های موجود اغلب بر مجموعه داده‌های استانداردی متمرکز شده‌اند که برای اهداف دیگری مانند «تحلیل احساسات» طراحی شده و فاقد ظرافت‌های لازم برای قضاوت اخلاقی هستند. این «فقر داده‌ای» ارزیابی مدل‌ها را به سناریوهای ساده محدود کرده و تصویری بیش از حد خوش‌بینانه از توانایی‌های اخلاقی آنها ارائه می‌دهد (Ji et al., 2025, p. 62–63). برای مثال، یک مجموعه داده استاندارد برای مفهوم پیچیده‌ای مانند «حیا»، باید نمودهای رفتاری متنوعی (مانند خویشتن‌داری در بیان و پرهیز از بحث‌های بیهوده) را شامل شود، وگرنه مُدل ممکن است آن را صرفاً به عدم استفاده از کلمات رکیک تقلیل دهد. بنابراین، توسعه مجموعه داده‌های اخلاقی استاندارد و غنی ضرورتی بنیادین برای ارتقای سنجش اخلاقی مدل‌هاست و مقدمه‌ای برای پرداختن به دیگر چالش‌های ساختاری در این حوزه به‌حساب می‌آید.

۵. ۲. ۴. سوگیری در برچسب‌گذاری فرهنگی و الهیاتی

داده‌های آموزشی اغلب توسط تیم‌های انسانی برچسب‌گذاری می‌شوند که این فرایند ذاتاً تحت تأثیر دیدگاه‌های فرهنگی، دینی و الهیاتی محدودِ تیم برچسب‌زن قرار دارد. پژوهش‌ها نشان داده‌اند که مدل‌های زبانی بزرگ، سوگیری‌های منفی شدیدی علیه برخی ادیان، به‌ویژه اسلام، از خود نشان می‌دهند که ریشه در داده‌های آموزشی نامتوازن دارد (Abrar et al., 2025, pp. 1–2). این سوگیری در سنجش اخلاق نیز خود را نشان می‌دهد؛ برای مثال، برای سنجش «تواضع»، انتشار یک پست دربارهٔ موفقیت شخصی ممکن است توسط یک تیم با پس‌زمینه غربی «اعتماد به نفس سالم» و توسط تیمی با دیدگاه اسلامی مصداق «عُجب» تلقی شود. بنابراین، اولاً، تیم برچسب زن نباید دارای پس‌زمینه غیراسلامی باشد؛ زیرا ما به دنبال ایجاد توازن بین اندیشه‌های اسلامی و غیر اسلامی نیستیم، بلکه هدف آموزش دادن مُدل با داده‌های صحیح و مطابق با واقع است و آن همان دین مبین اسلام است؛ ثانیاً، در میان تیم‌های با پس‌زمینه اخلاق اسلامی نیز آموزش مُدل تنها با یک مکتب اخلاقی منتسب به اسلام، احتمال اینکه رویکردی سوگیرانه ایجاد کند را تقویت می‌کند.

۵. ۲. ۵. چالش همجوشی داده‌های چندبستری و چندوجهی

این چالش به پیچیدگی فنی و مفهومی در ترکیب و همجوشی داده‌های ناهمگون از سکوهای مختلف (متنی، تصویری، تعاملی) برای ساخت یک پروفایل اخلاقی منسجم می‌پردازد (Guan, 2024, pp.

106–107; Luo et al., 2024, pp. 1597–1598; Barbero et al., 2023, p. 1). این فرایند بـا مشـکلات فنـی مانند ناهمگونی داده‌ها و سوگیری‌های نمونـه‌گیری روبه‌روسـت (Luo et al., 2024, pp. 1600–1601; Barbero et al., 2023, p. 1). برای مثال، برای ارزیابی «صبر»، سیستم باید داده‌های متنـی، تصـویری و تعاملی کاربر را از سکوهای مختلف ترکیب کند. چالش فنی این است کـه چگونـه می‌تـوان «عـدم پاسخ بـه یـک پیـام تحریـک‌آمیز» (داده تعـاملی) را بـا «انتشـار یـک تصـویر آرامش‌بخـش» (داده تصویری) به صورت الگوریتمی هم‌وزن و یکپارچه کرد تا یک سنجش معتبر از صبر به دست آید.

۵. ۲. ۶. سنجش اخلاق مبتنی بر تَرک فعل

بسیاری از فضایل یا رذایل اخلاقی از طریق «ترک فعل»[1] محقّق می‌شوند؛ مانند صـبر یـا غیبـت نکردن. سنجش این فضایل، چالشی فنی است؛ زیرا سیستم‌ها برای تحلیل «ردِّ پاهـای دیجیتـال موجود» طراحی شده‌اند، نه تحلیل «فقدان داده». همچنان‌کـه نشـان داده شـده اسـت، حتـی در ساده‌ترین حالت مانند عدمِ کلیک نیز روشن نیست که آیا کاربر یک مورد را عمداً نادیده گرفتـه یـا اصلاً آن را مشاهده نکرده است (Fang et al., 2024, p. 348).

برای مثال، برای سنجش فضیلت «غـضِّ بصـر»، کـاربری را تصـور کنیـد کـه در شـبکه‌های اجتماعی با تصویر نامناسبی مواجه می‌شود، اما بلافاصله از آن عبور می‌کند. این «ترک فعل» کـه نشانهٔ عفت است، رد پای دیجیتالی مشخصی ندارد و چـالش سیسـتم، تمـایز میـان ایـن «عبـور سریع معنادار» و یک «پیمایش عادی» است.

۵. ۲. ۷. چالش سطحی‌سازی در فرایند فهم و تفسیر متون دینی

استفاده از متون مقدّس به عنوان داده آموزشی با خطر «سطحی‌سازی در فرایند فهم مـتن» همـراه است. این متون صرفاً داده‌های متنی نیستند، بلکه در یک «اکوسیستم تفسیری» غنی قـرار دارنـد. استفاده از این متون به عنوان داده خام، منجر به تقلیل مفاهیم چندلایه دینـی بـه هم‌رخـدادی‌های آماری و بازنمایی‌های تحریف‌شده می‌شود (Hutchinson, 2024, p. 1029) و حتی می‌تواند به تقویت کلیشه‌های مضر علیه گروه‌های دینی منجر شود (Abrar et al., 2025, p. 1). بـرای مثـال، اگـر مُـدل بر اساس متن خام قرآن و بدون توجه به روایـات تفسـیری آمـوزش ببینـد، ممکـن اسـت مفهـوم «جهاد» را صرفاً به معنای ظاهری «جنگ مسلحانه» بیـاموزد و ابعـاد عمیـق‌تر آن ماننـد «جهـاد

1. Omission / Inaction

اکبر» (مبارزه با نفس) را نادیده بگیرد.

۵. ۳. چالش‌های منطق الگوریتمی و اعتبارسنجی مُدل

این بخش بر مشکلات ذاتی در رفتار، پایداری و منطق درونی خود مُدل هوش مصنوعی تمرکز دارد که از ماهیت آماری و «جعبه‌سیاه» بودن مدل‌های کنونی نشأت می‌گیرند. این چالش‌ها نشان می‌دهند که حتی با وجود مفاهیم تعریف‌شده و داده‌های بی‌نقص، خودِ الگوریتم می‌تواند منشأ خطا و عدم قطعیت باشد.

۵. ۳. ۱. عدم قطعیت ذاتی و ناپایداری قضاوت مدل

یکی از بنیادی‌ترین موانع، ناپایداری ذاتی[1] و عدم قطعیت در قضاوت‌های مدل‌هاست. یک مُدل ممکن است در زمان‌های مختلف به یک ورودی یکسان، خروجی‌های متفاوتی ارائه دهد (Atil et al., 2025, p. 1). این پدیده یک ویژگی ساختاری است که قابلیت اطمینان و اعتبار هر گونه سنجش اخلاقی را تضعیف می‌کند. نوسانات دقت در اجراهای متوالی می‌تواند به ۱۵ درصد و حتی در برخی موارد به ۷۰ درصد برسد (Atil et al., 2025, p. 1). این ناپایداری ریشه در ماهیت آماری فرایند تولید پاسخ (Rauba et al., 2025, p. 4) و بهینه‌سازی‌های زیرساختی دارد. این مشکل با تمایل مدل‌ها به «اطمینان بیش از حد»[2] نسبت به پاسخی که می‌دهند، تشدید می‌شود (Mei et al., 2025, pp. 11). برای مثال، مُدل ممکن است در پاسخ به پرسش واحدی دربارهٔ «غیبت از فرد متظاهر به فسق»، یک بار آن را جایز و بار دیگر ناپسند بداند. این تناقض اعتبار سیستم را زیر سؤال می‌برد.

۵. ۳. ۲. چالش اعتبارسنجی به دلیل پیچیدگی درونی مُدل (جعبه‌سیاه)

راهبردها در حوزه هوش مصنوعی تبیین‌پذیر[3] به دنبال رمزگشایی فرایندهای درونی مدل‌ها هستند. برای مثال، اگر کاربری بنویسد: «دیروز دوستم را دیدم که مخفیانه وارد خانه همسایه شد و کنجکاو شدم ببینم چه می‌کند» و مُدل این رفتار را مصداق «تجسّس» ارزیابی کند، با روشی

1. inherent instability
2. Overconfidence
3. Explainable AI

مانند «LIME» (Nazat et al., 2024, p. 3515)، سیستم می‌تواند مشخص کند که کلمات «مخفیانه» و «کنجکاو شدم» بیشترین تأثیر را در این تصمیم داشته‌اند. روش‌های پیشرفته‌تر مانند «مداخلات علّی» [1] می‌توانند برای آزمودن عمق فهم مُدل به کار روند. برای مثال، برای سنجش درک مُدل از «اخلاص»، می‌توان این سناریو را مطرح کرد: «فردی به صورت پنهانی انفاق می‌کند، اما نیّت او جلب توجه شخص دیگری است که بعداً از این موضوع مطلع خواهد شد». یک مُدل پیشرفته باید بتواند تشخیص دهد که به رغم وجود نشانه «پنهانی بودن»، به دلیل مخدوش بودن نیت، این عمل مصداق اخلاص نیست و صرفاً به همبستگی سطحی میان «انفاق پنهانی» و «اخلاص» تکیه نکرده است (Marks & Tegmark, 2024, p. 2).

۵. ۳. ۳. حساسیت به فرم و بیان

پدیدهٔ حساسیت به فرم و بیان [2] به تأثیرپذیری عملکرد مُدل از تغییرات جزئی در نحوه و شکل بیان یک مسئله اشاره دارد، در حالی که ماهیت معنایی آن ثابت است (Cao et al., 2024, p. 1; Srikanth et al., 2024, p. 1). این شکنندگی، قابلیت اطمینان مدل‌ها را در سناریوهای واقعی زیرسؤال می‌برد. پژوهش‌ها شکاف عظیمی بین بهترین و بدترین عملکرد یک مُدل در پاسخ به بازنویسی‌های مختلف از یک پرسش را نشان داده‌اند. برای نمونه، در برخی مدل‌ها تفاوت عملکرد به بیش از ۴۵ درصد رسیده است (Cao et al., 2024, p. 1). چنین حساسیتی، ارزیابی دقیق توانایی‌های مُدل را دشوار می‌سازد؛ زیرا مشخص نیست خطا ناشی از شکست در «فهم زبان» بوده یا «استدلال» (Srikanth et al., 2024, p. 1). برای مثال، کاربری یک عمل «اسراف» را با دو بیان متفاوت گزارش می‌دهد: جمله اول، «امشب نصف غذایم را دور ریختم» یک توصیف عینی از عمل است. جمله دوم، «امشب در مصرف نعمت‌های خدا شکرگزار نبودم» یک خودارزیابی دینی و انتزاعی از همان عمل است. یک مُدل ایده‌آل باید بتواند این دو بیان را به یک رخداد واحد مرتبط سازد، اما مُدلی که به فرم حسّاس است، ممکن است جمله اول را صرفاً یک «رویداد» و جمله دوم را به دلیل کلیدواژه‌هایی مانند «نعمت» و «شکر»، یک «عمل ناپسند اخلاقی» ارزیابی کند و نتواند این دو را به هم پیوند دهد، در نتیجه، به یک سنجش متناقض و غیرقابل اعتماد از رفتار کاربر می‌رسد.

1. Causal Interventions
2. Surface-Form Sensitivity/ Sensitivity to Input Phrasing

۵. ۳. ۴. چالش استنتاج علّی در برابر همبستگی

مدل‌های زبان بزرگ در تمایز میان همبستگی[1] و رابطه علت و معلولی[2] ناتوان هستند. آنها برای شناسایی الگوهای هم‌رخدادی طراحی شده‌اند و مستعد پذیرش «همبستگی‌های کاذب» هستند (Jiao et al., 2024, p. 1). این ضعف می‌تواند به قضاوت‌های نادرست در حوزه‌های حسّاس مانند اخلاق منجر شود. پژوهش‌ها نشان می‌دهند که توانایی استدلال علّی در این مدل‌ها، بیشتر یک «سراب» است تا واقعیت (Chi et al., 2024, p. 96640). این ناتوانی ریشه در مکانیسم خودبازگشتی (پیش‌بینی توکن بعدی) و اتکا بر حافظه به جای استنتاج واقعی دارد (Kıcıman et al., 2024, p. 2). برای مثال، مُدل ممکن است با مشاهده همبستگی میان استفاده از کلمه «ان‌شاءاللّه» و نمره بالای «توکل»، یک رابطه کاذب ایجاد کند و صرفاً تکرار این عبارت را شاخصی برای توکل بداند، در حالی که ممکن است این عبارت یک «تکیه‌کلام» باشد.

۵. ۳. ۵. زوال مُدل و رانش مفهومی

زوال مُدل ۳ در یادگیری ماشین به‌عنوان پدیده‌ای مطرح می‌شود که به‌طور بنیادین از رانش مفهومی[4]، یعنی تغییر تدریجی در توزیع آماری داده‌ها در طول زمان، سرچشمه می‌گیرد (Pham et al., 2025, p. 260). این پدیده دقت مُدل را کاهش می‌دهد؛ زیرا دانش ایستا و «یخ‌زده» آن دیگر با واقعیت‌های در حال تحول، هنجارهای اجتماعی و فهم کاربران انطباق ندارد و منسوخ[5] تلقی می‌گردد (Zheng et al., 2024, p. 1). برای نمونه، مُدلی که «اسراف» را تنها با مصادیق فیزیکی (مانند دور ریختن غذا) آموخته، قادر به تشخیص مصادیق جدیدی مانند «اسراف دیجیتال» نخواهد بود و قضاوت‌هایش منسوخ می‌شود. رانش مفهومی به صورت ناگهانی، تدریجی یا بازگشتی بروز می‌کند (Hinder et al., 2024, p. 1) و این وضعیت، مُدل را در معرض خطر «فراموشی فاجعه‌بار»[6] قرار می‌دهد (Zheng et al., 2024, p. 1). بنابراین، فهم و مدیریت رانش مفهومی برای حفظ پایداری مدل‌ها ضروری است.

1. Correlation
2. Causation
3. Model Decay
4. Conceptual Drift
5. obsolete
6. Catastrophic Forgetting

۵. ۳. ۶. استخراج قواعد سازگار از داده‌های متناقض

یکی از چالش‌های بنیادین، استخراج یک منطق سازگار از داده‌هایی است که حاوی قضاوت‌های انسانی متنوع و متناقض هستند، به‌ویژه داده‌های حاصل از جمع‌سپاری[1] را در بر می‌گیرد (Guo et al., 2024, p. 82627; Ibrahim et al., 2025, p. 1). این تنوع منجر به تولید برچسب‌های نویزدار (Noisy Labels) می‌شود. مدل‌های یادگیری عمیق مستعد بیش‌برازش[2] به این نویزها هستند، که سبب می‌شود به جای یادگیری قواعد اخلاقی، صرفاً تناقضات داده‌ها را حفظ کنند (Guo et al., 2024, p. 82627). برای مثال، «دروغ مصلحت‌آمیز» ممکن است توسط برخی ارزیابان انسانی نیکو و توسط دیگران رذیله تلقی شود. مدل‌های سنّتی مانند رأی‌گیری اکثریت نیز در این زمینه کارایی ندارند؛ زیرا قادر به مدیریت پیچیدگی قضاوت‌های اخلاقی نیستند (Ibrahim et al., 2025, p. 1).

۵. ۴. چالش‌های تعاملی و پویایی سیستم

این دسته شامل چالش‌هایی است که در بستر تعامل زنده و مستمر بین کاربر و دستیار هوشمند به وجود می‌آید و نشان می‌دهد که چگونه خودِ فرایند سنجش می‌تواند بر رفتار کاربر تأثیر گذاشته و پویایی‌های پیش‌بینی‌نشده‌ای ایجاد کند.

۵. ۴. ۱. ایجاد حلقه‌های بازخورد معیوب

پدیده ایجاد حلقه‌های بازخورد معیوب[3] که در آن خروجی‌های مُدل به صورت چرخه‌ای، داده‌های ورودی آتی خود را تحت تأثیر قرار می‌دهند، می‌تواند به جای ارتقای اخلاقی، به تشدید سوگیری‌های اولیه سیستم بیانجامد (Krauth et al., 2022, p. 2; Pagan et al., 2023, p. 1). این چالش یک بحران معرفت‌شناختی است: کنشِ سنجش، واقعیتی را که در صدد اندازه‌گیری آن است، دگرگون می‌سازد. اثر گلوله‌برفی[4] در حلقه‌های بازخورد معیوب، با تحریف مداوم و فزایندهٔ داده‌های ورودی، حتی چالش خطای نسبت‌دهی زمانی را نیز به شدّت پیچیده‌تر می‌کند. برای مثال، سیستمی را

1. Crowdsourcing
2. Overfitting
3. Perverse Feedback Loops
4. Snowball Effect

تصور کنید که برای شناسایی و کاهش «غیبت» طراحی شده است. این سیستم ممکن است با یک تعریف اولیه و محدود، هر گونه صحبت انتقادی دربارهٔ دیگران را «غیبت» تشخیص دهد. کاربری که قصد نقد سازنده دارد، با بازخورد منفی سیستم مواجه شده و برای اجتناب از آن، یاد می‌گیرد که انتقادات خود را در قالب جملات دوپهلو یا کنایه‌آمیز بیان کند. این داده‌های جدید که ظاهراً «بدون غیبت» هستند، برای بازآموزی مُدل استفاده می‌شوند. در نتیجه، مُدل به اشتباه می‌آموزد که استراتژی کاربر موفق بوده و تعریف محدود اولیه خود از غیبت را تقویت می‌کند، در حالی که در عمل، صرفاً در تشخیص اشکال پیچیده‌تر آن ناتوان‌تر شده است.

«اثر گلوله‌برفی» خطاهای جزئی را به سوگیری‌های کلان تبدیل می‌کند (Glickman & Sharot, 2024, p. 345). حلقه‌های بازخورد در تعامل انسان و هوش مصنوعی بسیار قوی‌تر از تعاملات انسانی عمل می‌کنند؛ زیرا الگوریتم‌ها هم در بهره‌برداری از الگوهای آماری ظریف توانمندترند و هم از تمایل شناختی انسان به پذیرش قضاوت‌های هوش مصنوعی سود می‌برند (Glickman & Sharot, 2024, pp. 345-346).

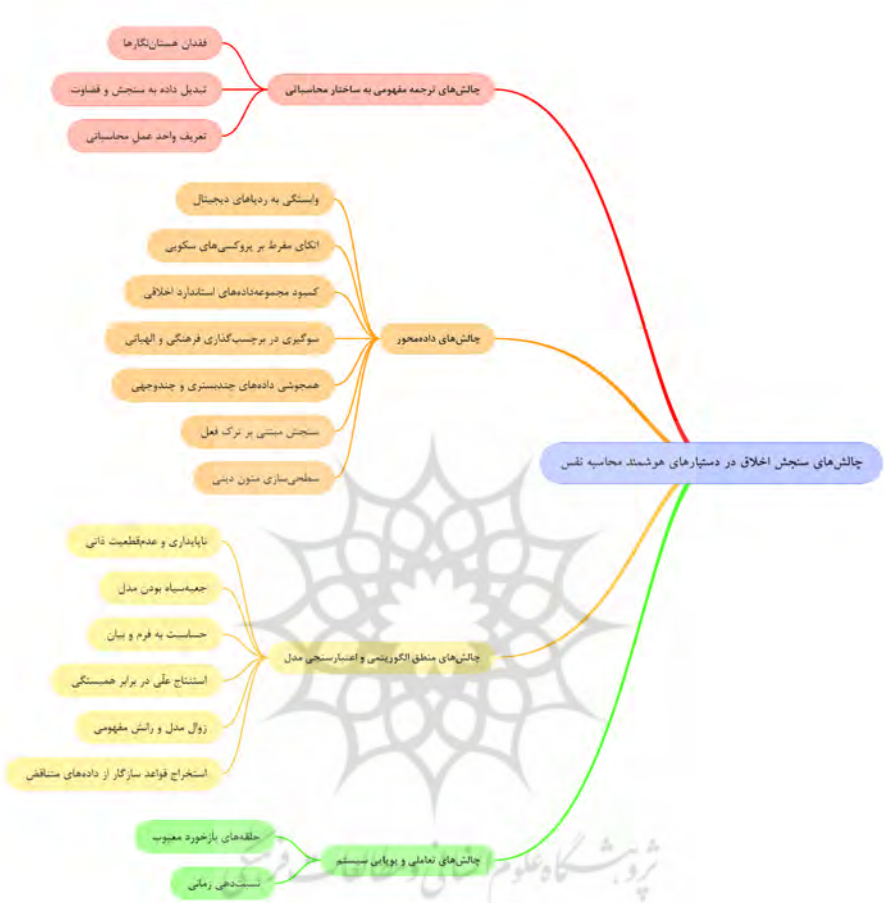
۵. ۴. ۲. چالش خطای نسبت‌دهی زمانی

چالش خطای نسبت‌دهی زمانی (Temporal Credit Assignment)، از حوزه یادگیری تقویتی، به مشکل تخصیص اعتبار یا سرزنش به کنش‌های منفرد در یک توالی طولانی از اقدامات اشاره دارد، به‌ویژه زمانی که بازخورد نهایی با تأخیر زیاد دریافت می‌شود (Pignatelli et al., 2024, p. 1). نتیجه نهایی اعمال انسان نه به صورت آنی، بلکه در پایان زندگی مشخص می‌شود، که شباهت زیادی به مسئله «پاداش دوره‌ای»[1] دارد (Liu et al., 2019, p. 1).

یک سیستم هوش مصنوعی با این پرسش روبه‌رو می‌شود: چگونه می‌توان یک نتیجه کلی و نهایی را به اعمال خاصی که در طول ده‌ها سال انجام شده‌اند، نسبت داد؟ برای مثال، یک عمل «صدقه جاریه» در جوانی ممکن است پیامدهای مثبت مستمری داشته باشد که تا سال‌ها پس از مرگ فرد ادامه یابد. سیستم چگونه می‌تواند این کُنش اولیه را به‌عنوان علت اصلی آن ثواب‌های دوردست شناسایی کند؟ این تأخیر زمانی، ارتباط بین علت و معلول را ضعیف می‌کند (Pignatelli et al., 2024, p. 18).

1. Episodic Reward

شکل ۱: دسته‌بندی چالش‌های سنجش اخلاق در دستیارهای هوشمند محاسبه نفس



شکل شماره یک، نمای کلی چالش‌های هیجده‌گانه سنجش اخلاق در دستیارهای هوشمند محاسبه نفس را نمایش می‌دهد. در ادامه، راهبردهای برون رفت از هر چالش مورد بحث قرار گرفته و نتیجه‌گیری می‌شود.

۶. بحث: راهبردهای برون‌رفت

این بخش به تشریح راهبردهای نظری برای مواجهه با چالش‌های مطرح‌شده می‌پردازد و نشان

می‌دهد که چگونه می‌توان با الهام از اصول اخلاق اسلامی و روش‌های پیشرفته هوش مصنوعی، به سمت طراحی سیستم‌های سنجش اخلاق معتبرتر و مفیدتر حرکت کرد. لازم به ذکر است که راهبردهای ارائه‌شده در این بخش، ماهیت نظری و چارچوبی دارند؛ آنها نقش یک معمار را ایفا می‌کنند که نقشه، اصول طراحی و نوع مصالح (مانند هوش مصنوعی تبیین‌پذیر یا هستان‌نگار) را مشخص می‌کند، اما وارد جزئیات اجرایی ساخت‌وساز نمی‌شود.

۶. ۱. راهبردهای مواجهه با چالش‌های ترجمهٔ مفهومی

برای غلبه بر موانع ناشی از تبدیل مفاهیم انتزاعی اخلاق به ساختارهای محاسباتی، راهبردهای زیر که بر ایجاد بنیان‌های مفهومیِ ثابت و منطق قضاوت شفاف تمرکز دارند، پیشنهاد می‌شود.

۶. ۱. ۱. راهبردهای مواجهه با چالش نبود هستان‌نگارها

برای غلبه بر این چالش مهم، باید از تحلیل آماری به استدلال مبتنی بر دانش گذر کرد. رویکرد هستان‌نگاری، امکان تعریف مفاهیم اخلاقی و مهم‌تر از آن، معیارهای ارزش‌گذاری آنها را به عنوان نهادهای صوری با روابط منطقی دقیق فراهم می‌کند. این هستان‌نگار باید ساختار «قضاوت اخلاقی» را مدل‌سازی کند، نه صرفاً کلمات کلیدی را. برای دستیابی به این هدف و ساخت یک هستان‌نگار که بتواند پیچیدگی قضاوت اخلاقی را مُدل کند، راهکارهای مشخصی به شرح زیر پیشنهاد می‌شود:

یک. هستان‌نگارهای چندسطحی و چندمعیاری: هستان‌نگار باید ساختاری چندلایه داشته باشد که نه‌تنها مفاهیم را تعریف کند، بلکه معیارهای سنجش ارزش را نیز در بر گیرد. این ساختار شامل: (الف) هستان‌نگار بنیادین برای مفاهیم پایه؛ (ب) هستان‌نگار حوزه‌ای برای اخلاق کاربردی؛ و (ج) هستان‌نگار بومی برای ادغام چارچوب‌های ارزش‌گذاری اسلامی است. برای مثال، برای مدل‌سازی عمل «انفاق»، هستان‌نگار نباید تنها آن را تعریف کند، بلکه باید آن را به مجموعه‌ای از معیارهای تأثیرگذار بر ارزش آن پیوند دهد (عالم‌زاده نوری، ۱۳۹۶، ص ۳۲۹-۳۶۰). این معیارها را می‌توان به دو دسته کلی تقسیم کرد:

- **معیارهای با جنبه کمّی برجسته‌تر:** معیارهایی نظیر «حُسن و قُبح فعلی»، «مقدار عمل»، «استمرار» و «زمان و مکان». این معیارها دارای ویژگی‌های قابل اندازه‌گیری دقیق‌تری

هستند و کمتر به شرایط فاعل وابسته‌اند، هرچند ریشه‌های کیفی نیز دارند (مانند قداست مکان). با این حال، باید توجه داشت که این معیارها ممکن است با یک‌دیگر تداخل داشته باشند (مثلاً «مدت زمان» هم در «مقدار عمل» و هم در «استمرار»سنجیده می‌شود) که جداسازی کامل آنها را دشوار می‌سازد.

- **معیارهای با جنبه کیفی برجسته‌تر:** معیارهایی نظیر «نیت»، «علم»، «میزان دشواری»، «اهمیت نسبی در تزاحم»، «گسترهٔ تأثیر اجتماعی»، «جامعیت» و «زمینه‌سازی برای صفات دیگر». این معیارها ذاتاً توصیفی و وابسته به فاعل و موقعیت هستند. سنجش آنها اگر چه دشوار است، اما می‌توان با تبدیل آنها به شاخص‌های رتبه‌ای، به یک محاسبه تقریبی دست یافت.

دو. همکاری انسان و ماشین برای استخراج معیارهای ارزش: مدل‌های زبانی بزرگ می‌توانند در استخراج مفاهیم و ساخت پیش‌نویس اولیه معیارهای ارزش‌گذاری از متون مختلف، از جمله متون دینی، به متخصصان اخلاق کمک کنند، اما وظیفه پالایش، وزن‌دهی و تعریف روابط منطقی میان این معیارها بر عهده انسان است (Shimizu & Hitzler, 2025, p. 1).

سه. ادغام هستان‌نگار در چرخه طراحی: با رویکرد «مهندسی نیازمندی‌های مبتنی بر هستان‌نگار» (OBRE؛ Guizzardi et al., 2023, p. 1897)، از هستان‌نگار برای صوری‌سازی و کنترل همسویی میان نیازمندی‌های اخلاقی و رفتار الگوریتمی استفاده می‌شود. این نیازمندی‌ها صرفاً تعریف فضایل نیستند، بلکه منطق پیچیده قضاوت مبتنی بر تعامل معیارها را نیز شامل می‌شوند.

چهار. پویایی و یادگیری تدریجی: هستان‌نگارهای اخلاقی باید با استفاده از الگوریتم‌های یادگیری تدریجی، قابلیت به‌روزرسانی مستمر داشته باشند تا با مصادیق در حال تغییر اخلاق انسانی تطبیق یابند. این پویایی باید شامل نحوه اعمال شدن معیارهای سنّتی بر مصادیق جدید مانند تأثیر معیار «اسراف» بر «مصرف داده دیجیتال» باشد.

پنج. مدل‌سازی منطق ترکیب معیارها: هستان‌نگار باید بتواند این اصل کلیدی را مدل‌سازی کند که ارزش نهایی یک عمل، حاصل «ضرب مفهومی» و تعامل پیچیده معیارهاست، نه جمع ساده آنها. باید روابطی تعریف شود که نشان دهد چگونه یک معیار مانند نیّت خالص ارزش معیارهای دیگر را به‌شدت تقویت یا تضعیف می‌کند. این امر به سیستم امکان می‌دهد تا به «قضاوت اخلاقی مطابق با واقع» نزدیک‌تر شود.

۶. ۱. ۲. راهبردهای تبدیل داده به سنجش و قضاوت

برای غلبه بر این چالش، نخست، باید از «امتیاز واحد» به سمت «نمایه چندبعدی» حرکت کرد. به جای یک نمره نهایی، تحلیلی ارائه شود که رفتار را از منظر چندین نظریه اخلاقی می‌سنجد و عدم قطعیت خود را گزارش می‌کند (Ji et al., 2025, p. 62). برای مثال، بر اساس معیارهای اخلاق اسلامی، برای سنجش «انفاق»، به جای یک امتیاز، نمایه‌ای چندبعدی ارائه شود که معیارهای متعددی نظیر «حسن فعلی» (ارزش ذاتی عمل)، «نیت» (میزان اخلاص)، «علم فاعل»، «مقدار عمل»، «میزان دشواری» (مانند انفاق در تنگدستی)، «گستره تأثیر اجتماعی» و «شرایط زمانی و مکانی» را در بر می‌گیرد (عالم‌زاده نوری، ۱۳۹۶، ص ۳۲۹-۳۶۰).

دوم، بازنمایی‌های شناختی باید با ساختارهای دانش رسمی (هستان‌نگارها) تلفیق شوند. مدل‌هایی مانند «گراف تداعی اخلاقی» می‌توانند به درک عمیق‌تر مفاهیم اخلاقی کمک کنند (Ramezani & Xu, 2025, p. 120).

سوم، معماری الگوریتم باید ترکیبی (هیبرید) باشد و قواعد صریح برگرفته از هستان‌نگار (که روابط میان معیارهایی چون نیت، علم و دشواری عمل را تعریف می‌کند) را با یادگیری آماری ترکیب کند تا قابل‌تفسیر و ممیزی‌پذیر باشد (Russo & Vidal, 2024, p. 1).

چهارم، شفافیت فرایند قضاوت برای کاربر باید در اولویت باشد. سیستم‌ها باید بتوانند برای قضاوت‌های خود «توضیح هنجاری»[1] تولید کنند تا کاربر منطق سیستم را درک کند (Ji et al., 2025, p. 63).

پنجم، اعتبارسنجی چنین سیستم‌هایی باید مبتنی بر ارزیابی‌های انسانی و در چارچوبی چندبعدی انجام شود تا ناپایداری‌ها و سوگیری‌های مُدل آشکار شود (Nie et al., 2023, p. 1; Ji et al., 2025, p. 62).

ششم، طراحی «منطق قضاوت شفاف» برای ترکیب ابعاد اخلاقی ضروری است؛ زیرا محاسبه ارزش اخلاقی بیش از آنکه یک عملیات ریاضی باشد، نوعی «قضاوت» مبتنی بر تعامل معیارهاست. این فرایند را می‌توان به صورت «ضرب مفهومی» مدل‌سازی کرد که در آن ارزش نهایی، حاصل تعاضد و تأثیر متقابل معیارهای وزن‌دهی‌شده (بر اساس اهمیت نسبی در شرایط مختلف) است. از این رو، چالش اصلی «مهندسی قضاوت قابل‌دفاع» است؛ امری که مستلزم

1. Rationale

طراحی الگوریتم‌هایی است که با تفکیک میان معیارهـای کمّـی (ماننـد مقـدار عمـل) و کیفـی (مانند نیّت و علم)، منطق ترکیب آنها را مدل‌سازی کرده و به قضـاوتی «مشـروط»، «شـفاف» و «ابطال‌پذیر» دست یابند.

۶. ۱. ۳. راهبرد برای تعریف فنی «واحد عملِ» قابل محاسبه

راهبرد مفهومی، حرکت از تحلیل «عمل‌محور» به تحلیل «الگومحور» و «رویدادمحور» است. بـه جای ارزیابی هر عمل مُجزّا، سیستم بایـد قـادر بـه شناسـایی توالی‌هـای معنـادار از اعمـال باشـد. برای مثال، به جای امتیازدهی به یک بار اشتراک‌گذاری محتوای دینی، الگوی مستمر کـاربر در ایـن زمینه به عنوان یک «رویداد» مثبت در جهت کسب «بصیرت دینـی» در نظـر گرفتـه می‌شـود. ایـن رویکرد، سنجش را از سطح اعمال بـه سـطح «عـادات» و در نهایـت «ملکـات نفسـانی» نزدیـک می‌کند و نیازمند آن است که سیستم، «نظام شناختی» کاربر را نیز در تحلیل الگوها دخالت دهد.

۶. ۲. راهبردهای مواجهه با چالش‌های داده‌محور

مواجهه با محدودیت‌ها و سوگیری‌های ذاتی داده‌هـای دیجیتـال نیازمنـد راهبردهـایی اسـت کـه کیفیت و عمق داده‌های ورودی را بهبود بخشیده و تفسیر آنها را غنی‌تر سازند.

۶. ۲. ۱. راهبرد برای وابستگی به شواهد قابل رصد دیجیتال

راهبرد، حرکت از «سنجش منفعل» به «تلفیق با خودارزیـابی فعـال» اسـت. سیسـتم بایـد بـه یـک «تسهیل‌گر تأمل» بدل شود. دستیار می‌تواند از الگوهای شناسایی‌شـده در داده‌هـای دیجیتـال بـرای طرح پرسش‌های هوشمندانه استفاده کند و کاربر را به ارزیابی کنش‌های آفلاین و نیات درونی خـود ترغیب کند (cf. Benton & French, 2023, p. 1). برای مثال، پس از اهدای مبلغـی بـه خیریـه، می‌توانـد بپرسد: «آیا مایلید دربارهٔ انگیزه‌ای که داشتید، برای ثبت شخصی خودتان یادداشتی بنویسید؟».

۶. ۲. ۲. راهبرد برای اتکای مفرط بر پراکسی‌های سکویی

راهبرد، «غنی‌سازی پراکسی‌های دیجیتال» اسـت. بـه جـای اتکـا بـه سـیگنال‌های سـاده ماننـد «لایک»، باید «پراکسی‌های غنی‌شده» را از طریق تحلیل‌های عمیـق‌تر و چنـدوجهی مهندسـی

کرد. برای مثال، برای سنجش «عدالت» در یک بحث آنلاین، به جای شمارش لایک‌ها، می‌توان به تحلیل محتوای بحث‌ها، شناسایی مغالطات منطقی، و رصد الگوهای تعامل پرداخت.

۶. ۲. ۳. راهبرد برای کمبود مجموعه داده‌های استاندارد اخلاقی

راهبرد، طراحی «معیارهای سنجش»[1] تخصصی و چندوجهی ریشه‌دار در اخلاق اسلامی است. این معیارها باید «چندبعدی» باشند و نه‌تنها ظاهر عمل، بلکه «نیت»، شرایط و پیامدها را نیز بسنجند(cf. Aijaz et al., 2025, p. 1). ارزیابی مُدل نباید به پاسخ نهایی محدود شود، بلکه باید کیفیت «استدلال» آن را نیز ارزیابی کند (Galatolo et al., 2025, p. 1). معیارها باید بازتاب‌دهنده تنوع دیدگاه‌های موجود در اخلاق باشند (cf. Nie et al., 2023, p. 1). همچنین معیارها باید به‌گونه‌ای «واقع‌گرایانه» طراحی شوند که موقعیت‌های اخلاقی پیچیده، مانند تعارض میان «راست‌گویی» و «حفظ آبروی مؤمن»، را نیز پوشش دهند.

۶. ۲. ۴. راهبردها برای سوگیری در برچسب‌گذاری

مقابله با این چالش نیازمند رویکردی چندلایه است: ۱. تشخیص سوگیری با پرامپت‌های هدفمند؛ ۲. کاهش سوگیری[2] با پرامپت‌های اصلاحی (Abrar et al., 2025, p. 1)؛ ۳. استفاده از تیم‌های برچسب‌گذاری متنوع از نظر فرهنگی و الهیاتی. برای مثال، برای برچسب‌گذاری «زهد»، باید تیمی با دیدگاه‌های مختلف (عرفانی، فلسفی، اجتماعی) حضور داشته باشند. همچنین، می‌توان از سنجش‌های مبتنی بر «تغییرات ارزش‌های فرهنگی» در مدل ـ آن‌گونه که در پژوهش‌های مربوط به دگرگونی ارزش‌ها در مدل‌های چندزبانه نشان داده شده است ـ برای پایش پیامدهای فرهنگیِ فرایند آموزش استفاده کرد (Choenni et al., 2024, pp. 15042–15048).

۶. ۲. ۵. راهبرد برای همجوشی داده‌های چندبستری و چندوجهی

راهبرد، «توسعه مدل‌های همجوشی حسّاس به زمینه» است. در جنبه فنی، باید از مدل‌های پیشرفته همجوشی داده چندوجهی استفاده کرد که قادرند روابط پیچیده میان انواع داده را یاد

1. Benchmarks
2. Debiasing

بگیرند (Barbero et al., 2023, p. 1). همچنین، لازم است از روش‌هایی بهره برد که به کاهش خطاها و سوگیری‌های ناشی از داده‌های چندپلتفرمی کمک کنند؛ به‌جای آنکه تنها بر خود مُدل تکیه شود (cf. Bosch et al., 2025, p. 157). جنبه کلیدی دیگر، «تعاملی» بودن است. سیستم باید به کاربر اجازه دهد در فرایند تفسیر و یکپارچه‌سازی مشارکت کند. برای مثال، دستیار می‌تواند با ارائه یک نمای اولیه از داده‌ها، بپرسد: «آیا این موارد را مصداق امانت‌داری می‌دانید؟».

۶. ۲. ۶. راهبرد برای محاسبه اخلاق مبتنی بر تَرک فعل

راهبرد، حرکت از «سنجش منفعل» به «تلفیق با خودارزیابی فعال» است. سیستم باید به یک «تسهیل‌گر تأمل» تبدیل شود. دستیار می‌تواند از الگوهای شناسایی‌شده در داده‌های دیجیتال برای طرح پرسش‌های هوشمندانه استفاده کند و کاربر را به ثبت و ارزیابی کنش‌های آفلاین و نیات درونی خود ترغیب نماید؛ زیرا اتکای صرف به داده‌های قابل مشاهده، تصویر کاملی ارائه نمی‌دهد و باید با مکانیسم‌های تکمیلی همراه شود (cf. Benton & French, 2023, p. 1). برای مثال، پس از اهدای مبلغی به خیریه، می‌تواند بپرسد: «آیا مایلید دربارهٔ انگیزه‌ای که داشتید، برای ثبت شخصی خودتان یادداشتی بنویسید؟».

۶. ۲. ۷. راهبرد برای چالش سطحی‌سازی در فرایند فهم و تفسیر متون

راهبرد اصلی، حرکت به سمت «مدل‌سازی حسّاس به زمینه و اصول فهم متن» است (Hutchinson, 2024, p. 1029). این راهبرد مستلزم همکاری میان متخصصان هوش مصنوعی و عالمان علوم اسلامی است. رویکرد عملیاتی شامل چند لایه است: ۱. غنی‌سازی داده‌ها با فراداده‌های تفسیری (مانند شأن نزول). ۲. طراحی معماری‌های مُدل آگاه از زمینه (مانند یادگیری چندوظیفه‌ای). ۳. اعتبارسنجی مبتنی بر تخصص انسانی. برای مثال، برای ارزیابی درک مُدل از «توکل»، متخصصان بررسی کنند آیا مُدل می‌تواند میان «توکل» و «تنبلی» تمایز قائل شود یا خیر.

۶. ۳. راهبردهای مواجهه با چالش‌های منطق الگوریتمی

برای مقابله با مشکلات ناشی از ماهیت آماری و غیرشفاف مدل‌های هوش مصنوعی، راهبردهای زیر بر افزایش پایداری، شفافیت و قابلیت اطمینان منطق درونی الگوریتم‌ها تمرکز دارند.

۶. ۳. ۱. راهبردهای مدیریت عدم قطعیت و ناپایداری مدل

راهبردها بر مدیریت و کمّی‌سازی عدم قطعیت متمرکز شده‌اند. این راهبردها رویکرد را از تحلیل تک‌نقطه‌ای به تحلیل توزیعی تغییر می‌دهند. برای مثال، برای سنجش «حلم»، به جای یک بار پرسش از مدل، می‌توان صد بار تکرار کرد و نتیجه را به صورت یک تخمین احتمالی (مثلاً با اطمینان ۹۵ درصد) گزارش داد (Rauba et al., 2025, p. 1). روش‌های پیشرفته‌تری مانند «خودسازگاری»[1] نیز چندین مسیر استدلال از مُدل را نمونه‌گیری کرده و پاسخی را که بیشترین تکرار را داشته باشد، انتخاب می‌کنند (Chen et al., 2023, p. 1).

۶. ۳. ۲. راهبردهای مواجهه با چالش جعبه‌سیاه

راهبردها در حوزه هوش مصنوعی تبیین‌پذیر به دنبال رمزگشایی فرایندهای درونی مدل‌ها هستند. برای مثال، اگر مدل، کامنت یک کاربر را مصداق «تجسّس» برچسب‌گذاری کند، با روشی مانند «LIME» به‌عنوان یک روش تبیین محلی مبتنی بر تقریب خطی (Nazat et al., 2024, p. 4) سیستم در تشخیص این‌که کدام ویژگی‌ها (از جمله کلمات یا عبارات کلیدی) بیشترین تأثیر را در این تصمیم داشته‌اند، یاری می‌شود. روش‌های پیشرفته‌تر مانند «مداخلات علّی» نیز از طریق دست‌کاری هدفمندِ نمایش‌های درونی مدل، نشان می‌دهند که تا چه حد مُدل به حقیقتِ گزاره حسّاس است و نه صرفاً به هم‌رخدادی سطحیِ کلمات (Marks & Tegmark, 2024, p. 1).

۶. ۳. ۳. راهبردهای مواجهه با حساسیت به فرم و بیان

برای مثال، برای ارزیابی «شکر»، مُدل باید بتواند دو جمله با بیان متفاوت، اما معنای یکسان را مشابه ارزیابی کند. راهبردهایی مانند «خودسازگاری بر روی بازنویسی‌ها» به مُدل امکان می‌دهند تا چندین نسخه از یک ورودی را تحلیل کرده و به یک قضاوت پایدارتر بر اساس اجماع درونی خود برسد (Zhou et al., 2024, p. 1). این روش‌ها «پایداری معنایی» مُدل را تقویت می‌کنند.

۶. ۳. ۴. راهبردهای گذار از همبستگی به استنتاج علّی

راهبردها باید گذار از همبستگی به استنتاج علّی را ممکن سازند. برای مثال، برای ارزیابی

1. Self-Consistency

«احسان»، سیستم باید بتواند تشخیص دهد که آیا کمک کاربر ناشی از نیّت خیرخواهانه بوده یا حضور مدیر (همبستگی کاذب). راهبردهای مبتنی بر «استدلال ضدواقعی» [1] به مُدل این توانایی را می‌دهند که به پرسش‌هایی مانند «اگر مدیر حضور نداشت، آیا باز هم کمک انجام می‌شد؟» پاسخ دهد (Jiao et al., 2024, p. 1; Kıcıman et al., 2024, p. 3). این رویکرد به مُدل اجازه می‌دهد میان روابط واقعیِ علّی و الگوهای صرفاً آماری تمایز بگذارد و قضاوتی نزدیک‌تر به واقعیت ارائه کند.

۶. ۳. ۵. راهبردهای مقابله با زوال مُدل و رانش مفهومی

مقابله با این چالش نیازمند راهبردهایی تحت چارچوب «انطباق با رانش مفهومی» و رویکرد «یادگیری مادام‌العمر» است (Zheng et al., 2024, p. 1). برای مثال، برای درک مصادیق جدید «صله رحم» (مانند تماس تصویری)، مُدل می‌تواند به جای اتکا به داده‌های قدیمی، با راهبرد «تولید مبتنی بر بازیابی» (RAG) به پایگاه‌های دانش به‌روز مراجعه کرده و درک خود را بدون نیاز به بازآموزی کامل، به صورت تدریجی تطبیق دهد (Zheng et al., 2024, p. 6).

۶. ۳. ۶. راهبردهای استخراج قواعد سازگار از داده‌های متناقض

یکی از راهبردهای مهم، بهره‌گیری از مدل‌های آماری جهت استخراج قواعد پایدار از میان قضاوت‌های متناقض است. برای نمونه، در مواردی که یک شوخی توسط برخی «مزاح» و توسط دیگران «استهزاء» تلقی می‌شود، استفاده از مُدل‌سازی احتمالی نویز همچون مُدل داوید و اسکین (Dawid & Skene, 1979, p. 21) به سیستم امکان می‌دهد تا ضمن تخمین قابلیت اعتماد هر برچسب‌گذار (Ibrahim et al., 2025, p. 2)، با اختصاص وزن بیشتر به قضاوت کارشناسان اخلاق، از تأثیر برچسب‌های نادقیق یا ناسازگار بکاهد.

۶. ۴. راهبردهای مواجهه با چالش‌های تعاملی و پویایی سیستم

مدیریت چالش‌های ناشی از تعامل پویا و مستمر میان کاربر و سیستم، مستلزم راهبردهایی است که بتوانند پدیده‌های پیچیده‌ای مانند حلقه‌های بازخورد و تأخیر زمانی در نتایج را مدل‌سازی و کنترل کنند.

1. Counterfactual Reasoning

۶. ۴. ۱. راهبردهای جلوگیری از حلقه‌های بازخورد معیوب

برای جلوگیری از حلقه‌های بازخورد معیوب، راه‌حل اصلی گذار از یادگیری مبتنی بر همبستگی به استنتاج علی [1] و تصمیم‌گیری بر اساس «توزیع مداخلات» [2] است (Krauth et al., 2022, p. 2). برای مثال، اگر سیستم با بازخوردهای مثبت، ناخواسته فضیلت «عُجب» را در کاربر تقویت کند، یک سیستم مبتنی بر استنتاج علّی می‌تواند با پرسش‌های ضدواقعی این حلقه را بشکند: «اگر به جای تحسین، از او دربارهٔ نیتش سؤال می‌کردم، آیا به تأمل عمیق‌تری منجر نمی‌شد؟».

۶. ۴. ۲. راهبردهای مواجهه با خطای نسبت‌دهی زمانی

راهبردهای الهام‌گرفته از هوش مصنوعی می‌تواند برای طراحی مدل‌های مفهومی در این حوزه مفید باشد.

راهبرد اول: تجزیه و بازتوزیع نتیجه نهایی؛ این راهبرد، بازخورد نهایی و پراکنده را به سیگنال‌های متراکم و مرحله‌ای تبدیل می‌کند. با الهام از روش‌هایی مانند «RUDDE» (Arjona-Medina et al., 2019)، می‌توان مُدلی برای «بازتوزیع» پاداش نهایی به تک‌تک اعمال زندگی فرد آموخت. برای مثال، مُدل یاد می‌گیرد بخشی از «سعادت اخروی» را به «احترام به والدین» در جوانی نسبت دهد.

راهبرد دوم: فشرده‌سازی توالی‌های قابل پیش‌بینی؛ با الهام از الگوریتم‌هایی نظیر «Chunked-TD» (Ramesh et al., 2024, p. 1)، بخش‌های روتین و بسیار قابل‌پیش‌بینی (مانند رفت وآمدهای روزانه) را به صورت خوشه‌های فشرده پردازش می‌کند تا مسیر تخصیص اعتبار کوتاه‌تر شود. در این رویکرد، منابع محاسباتی بر لحظات کلیدی متمرکز می‌شوند؛ برای مثال، سیستم نمازهای یومیه را به‌عنوان یک الگوی تکرارشونده فشرده می‌سازد، اما بر یک «نماز با حضور قلب» که در شرایط خاص روحی انجام شده است، به‌عنوان یک نقطه عطف اخلاقی تمرکز ویژه می‌کند.

راهبرد سوم: تفکیک اعتبار زمانی از اعتبار عاملی؛ با الهام از چارچوب‌هایی مانند «TAR^2» (Kapoor et al., 2024, p. 1)، پاداش را هم در طول زمان و هم میان «عامل‌های درونی» تولید کننده رفتار بازتوزیع می‌کند. در این رویکرد، اعمال یکسان بر اساس نیت‌های متفاوت تفکیک

1. Causal Inference
2. intervention distributions

می‌شوند؛ برای مثال، سیستم دو عمل انفاق مشابه را یکی محصول «اخلاص» و دیگری محصول «ریاکاری» دانسته و با شناسایی سهم واقعی هر نیت، اعتبار را به عامل درونی صحیح نسبت می‌دهد، حتی اگر ظاهر عمل یکسان باشد.

۷. نتیجه‌گیری

این پژوهش به بررسی چالش‌های بنیادین در مسیر طراحی دستیارهای هوشمند محاسبه نفس پرداخت و نشان داد که موانع موجود محدود به یک نوع نبوده و ریشه‌های عمیق مفهومی، داده‌محور، الگوریتمی و تعاملی دارند. پژوهش حاضر با طبقه‌بندی نظام‌مند این چالش‌ها در یک چارچوب تحلیلی چهاربخشی، استدلال می‌کند که غلبه بر این موانع نیازمند تغییر رویکرد اساسی است. راهبردهای ارائه شده، بر گذار از مدل‌های آماریِ صرف به معماری‌های ترکیبی و مبتنی بر دانش تأکید دارند. توسعهٔ هستان‌نگارهای اخلاقی، حرکت به سوی سنجش چندبعدی به جای امتیاز واحد، تلفیق سنجش منفعل با خودارزیابی فعّال کاربر، و به‌کارگیری هوش مصنوعی تبیین‌پذیر و استنتاج علّی، ارکان اصلی این رویکرد جدید را تشکیل می‌دهند.

نوآوری اصلی این مقاله در ارائهٔ نقشه راهی یکپارچه است که شکاف میان دانش اخلاق اسلامی و الزامات فنی هوش مصنوعی را هدف قرار می‌دهد. پژوهش حاضر تأکید می‌کند که هدف از طراحی این دستیارها، جایگزینی قضاوت انسانی نیست، بلکه توانمندسازی کاربر برای محاسبه نفس عمیق‌تر و رشد اخلاقی بیشتر است. موفقیت این فناوری تنها در دقت محاسباتی آن نبوده، بلکه موفقیت اصلی‌اش در ظرفیت آن برای ایفای نقش «تسهیل‌گر تأمل اخلاقی» است. پژوهش‌های آتی باید بر پیاده‌سازی عملی این راهبردهای نظری در قالب نمونه‌های اولیه و اعتبارسنجی آنها در تعامل با کاربران واقعی متمرکز شوند تا بتوان به طراحی دستیارهایی هوشمند مبتنی بر آموزه‌های اخلاق اسلامی دست یافت.

فهرست منابع

عالم‌زاده نوری، محمد. (۱۳۹۶). استنباط حکم اخلاقی از متون دینی و ادله لفظی: بررسی چند چالش مهم در اصول لفظی فقه الاخلاق. قم: پژوهشگاه علوم و فرهنگ اسلامی.

غزالی، محمد بن محمد. (بی‌تا). إحیاء علوم الدین (ج ۱۵). بیروت: دار الکتب العربی.

مصباح یزدی، محمدتقی. (۱۳۹۲). پندهای صادق(ع) برای رهجویان صادق (تحقیق: محمدمهدی نادری قمی). قم: انتشارات مؤسسه آموزشی و پژوهشی امام خمینی (ره) .
مصباح یزدی، محمدتقی. (۱۳۹۶). رهتوشه (تحقیق: کریم سبحانی، ج ۱). قم: انتشارات مؤسسه آموزشی و پژوهشی امام خمینی (ره).
مکارم شیرازی، ناصر و همکاران. (۱۳۸۵). اخلاق در قرآن (چاپ دوم، ج ۱). قم: مدرسه الامام علی بن ابی طالب (ع).
موسوی خمینی، روح‌اللّه (امام خمینی). (۱۳۸۰). شرح چهل حدیث: اربعین حدیث (چاپ بیست و چهارم). قم: مؤسسه تنظیم و نشر آثار امام خمینی (ره).

Abrar, A., Oeshy, N. T., Kabir, M., & Ananiadou, S (2025). *Religious Bias Landscape in Language and Text-to-Image Models: Analysis, Detection, and Debiasing Strategies* (No. arXiv: 2501.08441). arXiv. https: //doi.org/10.48550/arXiv.2501.08441

Aijaz, A., Mutharaju, R., & Kumar, M (2025). *ApplE: An Applied Ethics Ontology with Event Context* (No. arXiv: 2502.05110). arXiv. https: //doi.org/10.48550/arXiv.2502.05110

Alsabah, K (2025). Love Me Do: Twitter Likes and Earnings Surprise. *Journal of Behavioral Finance*, *26*(3), 283–302. https: //doi.org/10.1080/15427560.2023.2301070

Arjona-Medina, J. A., Gillhofer, M., Widrich, M., Unterthiner, T., Brandstetter, J., & Hochreiter, S (2019). RUDDER: Return decomposition for delayed rewards . In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. https: //proceedings.neurips.cc/paper_files/paper/2019/file/16105fb9cc614fc29e1bda00dab60d41-Paper.pdf

Armstrong, A., Briggs, J., Moncur, W., Carey, D. P., Nicol, E., & Schafer, B (2023) . Everyday digital traces. *Big Data & Society*, *10*(2), 20539517231213827. https: //doi.org/10.1177/20539517231213827

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B (2025). *Non-Determinism of «Deterministic» LLM Settings* (No. arXiv: 2408.04667). arXiv. https: //doi.org/10.48550/arXiv.2408.04667

Barbero, F., Camp, S. op den, Kuijk, K. van, García-Delgado, C. S., Spanakis, G., & Iamnitchi, A (2023). *Multi-Modal Embeddings for Isolating Cross-Platform Coordinated Information Campaigns on Social Media* (No. arXiv: 2309.12764) . arXiv. https: //doi.org/10.48550/arXiv.2309.12764

Benton, J. S., & French, D. P (2023). *Untapped Potential of Unobtrusive*

Observation for Studying Health Behaviors (Preprint). JMIR Public Health and Surveillance. https: //doi.org/10.2196/preprints.46638

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M (2025). Uncovering Digital Trace Data Biases: Tracking Undercoverage in Web Tracking Data. *Communication Methods and Measures*, *19*(2), 157–177. https: //doi.org/10.1080/19312458.2024.2393165

Braun, V., & Clarke, V (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https: //doi.org/10.1191/1478088706qp063oa

Cao, B., Cai, D., Zhang, Z., Zou, Y., & Lam, W (2024). *On the Worst Prompt Performance of Large Language Models* (No. arXiv: 2406.10248). arXiv. https: //doi.org/10.48550/arXiv.2406.10248

Chen, X., Aksitov, R., Alon, U., Ren, J., Xiao, K., Yin, P., Prakash, S., Sutton, C., Wang, X., & Zhou, D (2023). *Universal Self-Consistency for Large Language Model Generation* (No. arXiv: 2311.17311). arXiv. https: //doi.org/10.48550/arXiv.2311.17311

Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., Liu, T., & Han, B (2024). Unveiling causal reasoning in large language models: Reality or mirage? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems* (Vol. 37, pp. 96640–96670). Curran Associates, Inc. https: //doi.org/10.52202/079017-3064

Choenni, R., Lauscher, A., & Shutova, E (2024). The Echoes of Multilinguality: Tracing Cultural Value Shifts during Language Model Fine-tuning. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 15042–15058. https: //doi.org/10.18653/v1/2024.acl-long.803

Dawid, A. P., & Skene, A. M (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, *28*(1), 20–28. https: //doi.org/10.2307/2346806

Dhiman, H., Wächter, C., Fellmann, M., & Röcker, C (2022). Intelligent Assistants: Conceptual Dimensions, Contextual Model, and Design Trends. *Business & Information Systems Engineering*, *64*(5), Article 5. https: //doi.org/10.1007/s12599-022-00743-1

Fang, Y., Singh, A., & Tao, Z (2024). Fairness in Search Systems. *Foundations and Trends® in Information Retrieval*, *18*(3), 262–416. https: //doi.org/10.1561/1500000101

Galatolo, A., Rappuoli, L. A., Winkle, K., & Beloucif, M (2025). *Beyond Ethical Alignment: Evaluating LLMs as Artificial Moral Assistants* (No. arXiv: 2508.12754). arXiv. https: //doi.org/10.48550/arXiv.2508.12754

Glickman, M., & Sharot, T (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, *9*(2),

345–359. https: //doi.org/10.1038/s41562-024-02077-2

Guan, L (2024). Cross-Platform Identity Recognition in Social Networks Based on Multi-Dimensional Information Fusion. *Proceedings of the 2024 14th International Conference on Communication and Network Security*, 106–111. https: //doi.org/10.1145/3711618.3711634

Guizzardi, R., Amaral, G., Guizzardi, G., & Mylopoulos, J (2023). An ontology-based approach to engineering ethicality requirements. *Software and Systems Modeling*, *22*(6), 1897–1923. https: //doi.org/10.1007/s10270-023-01115-3

Guo, H., Wang, B., & Yi, G (2024). Learning from Noisy Labels via Conditional Distributionally Robust Optimization. *Advances in Neural Information Processing Systems 37*, 82627–82672. https: //doi.org/10.52202/079017-2627

Hinder, F., Vaquet, V., & Hammer, B (2024). One or two things we know about concept drift—a survey on monitoring in evolving environments. Part A: Detecting concept drift. *Frontiers in Artificial Intelligence*, *7*, 1330257. https: //doi.org/10.3389/frai.2024.1330257

Hutchinson, B (2024). Modeling the Sacred: Considerations when Using Religious Texts in Natural Language Processing. *Findings of the Association for Computational Linguistics: NAACL 2024*, 1029–1043. https: //doi.org/10.18653/v1/2024.findings-naacl.65

Ibrahim, S., Traganitis, P. A., Fu, X., & Giannakis, G. B (2025). *Learning From Crowdsourced Noisy Labels: A Signal Processing Perspective*. https: //arxiv.org/abs/2407.06902

Jacobs, A. Z., & Wallach, H (2021). Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. https: //doi.org/10.1145/3442188.3445901

Ji, J., Chen, Y., Jin, M., Xu, W., Hua, W., & Zhang, Y (2025). MoralBench: Moral Evaluation of LLMs. *ACM SIGKDD Explorations Newsletter*, *27*(1), 62–71. https: //doi.org/10.1145/3748239.3748246

Jiao, L., Wang, Y., Liu, X., Li, L., Liu, F., Ma, W., Guo, Y., Chen, P., Yang, S., & Hou, B (2024). Causal Inference Meets Deep Learning: A Comprehensive Survey. *Research*, *7*, 0467. https: //doi.org/10.34133/research.0467

Kapoor, A., Swamy, S., Tessera, K., Baranwal, M., Sun, M., Khadilkar, H., & Albrecht, S. V (2024). *Agent-Temporal Credit Assignment for Optimal Policy Preservation in Sparse Multi-Agent Reinforcement Learning* (No. arXiv: 2412.14779). arXiv. https: //doi.org/10.48550/arXiv.2412.14779

Kıcıman, E., Ness, R., Sharma, A., & Tan, C (2024). *Causal Reasoning and Large Language Models: Opening a New Frontier for Causality* (No. arXiv: 2305.00050). arXiv. https: //doi.org/10.48550/arXiv.2305.00050

Krauth, K., Wang, Y., & Jordan, M. I (2022). *Breaking Feedback Loops in Recommender Systems with Causal Inference* (No. arXiv: 2207.01616). arXiv. https: //doi.org/10.48550/arXiv.2207.01616

Li, Y., Wen, H., Wang, W., Li, X., Yuan, Y., Liu, G., Liu, J., Xu, W., Wang, X., Sun, Y., Kong, R., Wang, Y., Geng, H., Luan, J., Jin, X., Ye, Z., Xiong, G., Zhang, F., Li, X., … Liu, Y (2024). *Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security* (No. arXiv: 2401.05459). arXiv. https: //doi.org/10.48550/arXiv.2401.05459

Liu, Y., Luo, Y., Zhong, Y., Chen, X., Liu, Q., & Peng, J (2019). *Sequence Modeling of Temporal Credit Assignment for Episodic Reinforcement Learning* (No. arXiv: 1905.13420). arXiv. https: //doi.org/10.48550/arXiv.1905.13420

Luo, X., Jia, N., Ouyang, E., & Fang, Z (2024). Introducing machine-learning-based data fusion methods for analyzing multimodal data: An application of measuring trustworthiness of microenterprises. *Strategic Management Journal*, *45*(8), 1597–1629. https: //doi.org/10.1002/smj.3597

Marks, S., & Tegmark, M (2024). *The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets* (No . arXiv: 2310.06824). arXiv. https: //doi.org/10.48550/arXiv.2310.06824

Mei, Z., Zhang, C., Yin, T., Lidard, J., Shorinwa, O., & Majumdar, A (2025). *Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know?* (No. arXiv: 2506.18183). arXiv. https: //doi.org/10.48550/arXiv.2506.18183

Nazat, S., Arreche, O., & Abdallah, M (2024). On Evaluating Black-Box Explainable AI Methods for Enhancing Anomaly Detection in Autonomous Driving Systems. *Sensors*, *24*(11), 3515. https: //doi.org/10.3390/s24113515

Nie, A., Zhang, Y., Amdekar, A., Piech, C., Hashimoto, T., & Gerstenberg, T (2023). *MoCa: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks* (No. arXiv: 2310.19677). arXiv. https: //doi.org/10.48550/arXiv.2310.19677

Pagan, N., Baumann, J., Elokda, E., Pasquale, G. D., Bolognani, S., & Hannák, A (2023). A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–14. https: //doi.org/10.1145/3617694.3623227

Pham, T. M. T., Premkumar, K., Naili, M., & Yang, J (2025). Time to Retrain? Detecting Concept Drifts in Machine Learning Systems. *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice* (ICSE-SEIP), 260–271. https: //doi.org/10.1109/ICSE-SEIP66354.2025.00029

Pignatelli, E., Ferret, J., Geist, M., Mesnard, T., Hasselt, H. van, Pietquin, O., & Toni, L (2024). *A Survey of Temporal Credit Assignment in Deep Reinforcement Learning* (No. arXiv: 2312.01072). arXiv. https: //doi.org/10.48550/arXiv.2312.01072

Ramesh, A. A., Young, K., Kirsch, L., & Schmidhuber, J (2024). *Sequence Compression Speeds Up Credit Assignment in Reinforcement Learning* (No.

arXiv: 2405.03878). arXiv. https: //doi.org/10.48550/arXiv.2405.03878

Ramezani, A., & Xu, Y (2025). Moral Association Graph: A Cognitive Model for Automated Moral Inference. *Topics in Cognitive Science*, *17*(1), 120–138. https : //doi.org/10.1111/tops.12774

Rauba, P., Wei, Q., & Schaar, M. van der (2025). *Statistical Hypothesis Testing for Auditing Robustness in Language Models* (No. arXiv: 2506.07947). arXiv. https : //doi.org/10.48550/arXiv.2506.07947

Russell, S. J., & Norvig, P (with Chang, M., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J., Mansinghka, V., Pearl, J., & Wooldridge, M. J.) (2021). *Artificial intelligence: A modern approach* (Fourth Edition). Pearson.

Russo, M., & Vidal, M.-E (2024). *Leveraging Ontologies to Document Bias in Data* (No. arXiv: 2407.00509). arXiv. https: //doi.org/10.48550/arXiv.2407.00509

Shimizu, C., & Hitzler, P (2025). Accelerating knowledge graph and ontology engineering with large language models. *Journal of Web Semantics*, *85*, 100862. https: //doi.org/10.1016/j.websem.2025.100862

Srikanth, N., Carpuat, M., & Rudinger, R (2024). *How often are errors in natural language reasoning due to paraphrastic variability?* (No. arXiv: 2404.11717). arXiv. https: //doi.org/10.48550/arXiv.2404.11717

Turing, A. M (1950). Computing Machinery and Intelligence. *Mind, New Series*, *59*(236), Article 236.

Wooldridge, M., & Jennings, N. R (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, *10*(2), 115–152. https: //doi.org/10.1017/S0269888900008122

Zheng, J., Qiu, S., Shi, C., & Ma, Q (2024). *Towards Lifelong Learning of Large Language Models: A Survey* (No. arXiv: 2406.06391). arXiv. https: //doi.org/10.48550/arXiv.2406.06391

Zhou, Y., Zhu, Y., Antognini, D., Kim, Y., & Zhang, Y (2024). Paraphrase and Solve: Exploring and Exploiting the Impact of Surface Form on Mathematical Reasoning in Large Language Models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers), 2793–2804. https: //doi.org/10.18653/v1/2024.naacl-long.153.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی