

A Review and Formulation of Ethical Guidelines for the Development and Deployment of Artificial Intelligence Systems

Alireza Seghatoleslami*

PhD in Philosophy of Science; Assistant Professor; Information and Society Research Department; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran;
Email: Seghatoleslami@irandoc.ac.ir

Received: 19, Jul. 2025 | Accepted: 25, Aug. 2025

Abstract: The ethics of artificial intelligence is a relatively emerging field within the scope of applied ethics, which is influenced by the transformations driven by the development and deployment of AI systems and examines the ethical and social implications and issues associated with them. Although some of these consequences and ethical issues have been somewhat predictable and have enabled ethical analyses and investigations, it appears that as AI systems increasingly penetrate all dimensions and aspects of individual and social life, certain ambiguous and complex ethical challenges are also arising. Therefore, in the field of AI ethics studies, there does not actually exist a set of issues that are well-established or enjoy an acceptable level of comprehensiveness. Nonetheless, numerous efforts have been made by national and international organizations to collect and draft principles and ethical guidelines for AI in the past decade. Each AI ethics document generally consists of two parts: ethical principles and ethical guidelines. The ethical principles in these documents form the basis for formulating the ethical guidelines. The goal of this article is to introduce principles and develop ethical guidelines for AI ethics through the study of seven national and international documents. On this basis, the five integrative ethical principles of Floridi and Cowls were used as the foundation for compiling and examining AI ethical guidelines. These guidelines were reviewed in three stages. In the first stage, ethical guidelines were extracted from the aforementioned documents and listed within the framework of the five ethical principles. In the second stage, the collected guidelines were harmonized and refined. In the third stage, to evaluate the validity of the findings and the research outcome, the refined ethical guidelines were critiqued and reviewed by a focus group of experts under the five ethical principles to ensure the credibility of the selected ethical guidelines by achieving a relative consensus among the experts.

Keywords: AI Ethics, Ethical Principles, AI Ethics Document, Ethical Guidelines, Artificial Intelligence

* Corresponding Author

**Iranian Journal of
Information
Processing and
Management**

**Iranian Research Institute
for Information Science and Technology
(IranDoc)**

ISSN 2251-8223

eISSN 2251-8231

Indexed by SCOPUS, ISC, & LISTA

Vol. 41 | No. 2 | pp. 419-440

Winter 2026

<https://doi.org/10.22034/jipm.2025.2066388.2053>



بررسی و تدوین رهنمودهایی اخلاقی برای توسعه و استقرار سیستم‌های هوش مصنوعی

علیرضا ثقه‌الاسلامی

دکتری فلسفه علم؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات (ایرانداک)؛ تهران، ایران؛
پدیده‌آور رابط Seghatoleslami@irandoc.ac.ir



مقاله برای اصلاح به مدت ۶ روز نزد پدیده‌آوران بوده است.

پذیرش: ۱۴۰۴/۰۶/۰۳

دریافت: ۱۴۰۴/۰۴/۲۸

نشریه علمی | رتبه بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

شاپا (چاپی) ۲۲۵۱-۸۲۲۳

شاپا (الکترونیکی) ۲۳۰۱-۸۲۳۱

نمایه در SCOPUS، ISI، و LISTA

ijpm.irandoc.ac.ir

دوره ۴۱ | شماره ۲ | صص ۴۱۹-۴۴۰

زمستان ۱۴۰۴

<https://doi.org/10.22034/ijpm.2025.2066388.2053>



چکیده: اخلاق هوش مصنوعی حوزه مطالعاتی نسبتاً نوظهوری در گستره اخلاق کاربردی است که متأثر از تحولات ناشی از توسعه و استقرار سیستم‌های هوش مصنوعی به بررسی پیامدها و مسائل اخلاقی و اجتماعی مرتبط با آن می‌پردازد. اگرچه برخی از این پیامدها و مسائل اخلاقی تا حدی قابل پیش‌بینی بوده و امکان تحلیل‌ها و بررسی‌های اخلاقی را فراهم آورده است، اما به نظر می‌رسد با رشد و نفوذ روزافزون سیستم‌های هوش مصنوعی در تمامی ابعاد و شئون زندگی اجتماعی و فردی انسان‌ها، چالش‌های اخلاقی مبهم و پیچیده‌ای نیز در حال بروز است. این است که در حوزه مطالعاتی اخلاق هوش مصنوعی، مجموعه مسائلی که به خوبی تثبیت شده یا از جامعیت قابل قبولی برخوردار باشد، در واقع وجود ندارد. البته، تلاش‌های متعددی توسط سازمان‌های ملی و بین‌المللی در حوزه گردآوری و تدوین اصول و رهنمودهای اخلاقی هوش مصنوعی در یک دهه اخیر انجام شده است. هر سند اخلاق هوش مصنوعی به‌طور معمول، شامل دو بخش اصول اخلاقی و رهنمودهای اخلاقی است. در این سندها اصول اخلاقی، مبنای گردآوری و صورت‌بندی رهنمودهای اخلاقی است. هدف این مقاله معرفی اصول و تدوین رهنمودهای اخلاقی هوش مصنوعی با مطالعه هفت سند ملی و بین‌المللی است. بر این اساس، پنج اصل اخلاقی یکپارچه «فلوریدی و کولز» مبنای گردآوری و بررسی رهنمودهای اخلاقی هوش مصنوعی قرار گرفت. این رهنمودها در سه مرحله بررسی شد. در مرحله اول، رهنمودهای اخلاقی از اسناد مذکور استخراج و در چارچوب پنج اصل اخلاقی فهرست شدند. در مرحله دوم، رهنمودهای گردآوری‌شده، یکدست‌سازی و پالایش گردیدند. در مرحله سوم، برای ارزیابی روایی یافته‌ها و دستاورد پژوهش، رهنمودهای اخلاقی پالایش‌شده ذیل اصول اخلاقی پنج‌گانه، توسط گروهی کانونی از متخصصان، نقد و بررسی شد تا فرایند اعتباربخشی به رهنمودهای اخلاقی برگزیده با دستیابی

به اجماعی نسبی میان متخصصان انجام یابد.

کلیدواژه‌ها: اخلاق هوش مصنوعی، اصول اخلاقی، سند اخلاق هوش مصنوعی، رهنمودهای اخلاقی، هوش مصنوعی

۱. مقدمه

در دهه‌های آخر قرن بیستم، اولین ایده‌های طراحی و ساخت ماشین‌هایی با قابلیت شبیه‌سازی فرایندهای فعالیت‌های ذهنی انسان مطرح شد. این ماشین‌ها و ابزارها در واقع، سیستم‌های رایانشی هوشمندی هستند که با پردازش و مدل‌سازی کلان‌داده‌ها می‌توانند یاد بگیرند و استنتاج کنند، و در نهایت تصمیم‌گیری و عمل نمایند (بهراری ۱۳۹۸، ۱۶). با گسترش این سیستم‌های خودگردان هوشمند و پیشرفت علم رباتیک در چند دهه اخیر و ارتقای روزافزون حساسیت‌های اخلاقی نسبت به نفوذ هنجاری و تثبیت جایگاه سیستم‌های هوشمند در زندگی روزمره انسان‌ها، تحلیل‌گران و سیاست‌گذاران در ارتباط با چنین مباحثی، پیوسته درگیر سناریوهای افراطی میان دو-قطبی‌های خوش‌بینانه و بدبینانه نسبت به توسعه هوش مصنوعی و پیامدهای اخلاقی آن بوده‌اند. شاید بدین خاطر است که دهه‌های ۱۹۸۰ و ۱۹۹۰ به‌عنوان «زمستان هوش مصنوعی» شناخته می‌شود؛ دورانی که هوش مصنوعی به‌منزله تهدید و چالش دوران معاصر مطرح گردید. برخی از اندیشمندان گسترش و استفاده از سیستم‌های خودگردان هوشمند و فراگیر شدن استفاده از آن‌ها را تهدیدی مهلک برای جامعه بشری پیش‌بینی کردند.

با وجود این سناریوهای افراطی، انسان‌ها در سال‌های اخیر منافع بسیاری را در ابعاد نظری و عملی در توسعه و استقرار سیستم‌های هوش مصنوعی مشاهده کرده‌اند. از این‌رو، امروزه با فراگیر شدن استفاده از سیستم‌های خودمختار هوشمند و با توجه به دستاوردهای آن‌ها در حوزه‌های پزشکی، مهندسی، اقتصادی، حقوقی و موارد دیگر، بحث‌ها و بررسی‌های روان‌شناختی، جامعه‌شناختی، فلسفی، اخلاقی و حقوقی روش‌مندی حول این دستاوردها شکل گرفته است که نحوه مواجهه تحلیل‌گران و سیاست‌گذاران در توسعه هوش مصنوعی را تحت تأثیر قرار می‌دهد. هوش مصنوعی می‌تواند نقش مهمی در تسهیل، تدقیق و تسریع بسیاری از فعالیت‌ها در جوامع بشری ایفا کند؛ چرا که برای ارتقای کیفی فعالیت‌ها، به روش‌های هوشمندانه‌تری برای پردازش مقادیر بسیار زیاد داده، پایداری و کارآمدی نیاز داریم. با وجود این توصیف‌ها، هوش مصنوعی را باید

به‌عنوان یک فناوری متعارف تلقی کرد، نه به‌عنوان یک معجزه و نه به‌عنوان طاعون، بلکه به‌عنوان یکی از راه‌حل‌های بسیاری که نبوغ بشر توانسته است آن را ابداع کند. به همین دلیل است که مباحث اخلاقی مرتبط با توسعه هوش مصنوعی، همچون بسیاری از ملاحظات اخلاقی مرتبط با توسعه حوزه‌های متنوع علم و فناوری، پرسشی کاملاً انسانی بوده و از جایگاهی حیاتی برخوردار است (Floridi 2023, 2-3).

اخلاق هوش مصنوعی، حوزه مطالعاتی نسبتاً نوظهوری در گستره اخلاق کاربردی است که متأثر از تحولات ناشی از توسعه و استقرار سیستم‌های هوش مصنوعی به بررسی پیامدها و مسائل اخلاقی و اجتماعی مرتبط با آن می‌پردازد. اگرچه برخی از این پیامدها و مسائل اخلاقی تا حدی قابل پیش‌بینی بوده و امکان تحلیل‌ها و بررسی‌های اخلاقی را فراهم کرده است، اما به نظر می‌رسد با رشد و نفوذ روزافزون سیستم‌های هوش مصنوعی در تمامی ابعاد و شئون زندگی اجتماعی و فردی انسان‌ها، چالش‌های اخلاقی مبهم و پیچیده‌ای نیز در حال بروز است. این است که می‌توان گفت در حوزه مطالعاتی اخلاق هوش مصنوعی، مجموعه مسائلی که به‌خوبی تثبیت شده یا از جامعیت قابل قبولی برخوردار باشد، در واقع وجود ندارد (Müller 2020).

با گسترش و استفاده فراگیر از کامپیوتر و افزایش قدرت محاسباتی در ماشین‌های محاسبه‌گر، دهه ۱۹۵۰ تا ۱۹۷۰ را می‌توان سرآغاز «عصر طلایی هوش مصنوعی» تلقی کرد. از این‌رو، دولت‌ها و شرکت‌های خصوصی بزرگ سرمایه‌گذاری‌های کلانی برای تحقیقات نظری و کاربردی در زمینه هوش مصنوعی صرف کردند. اما با گذر زمان، برخی دانشمندان استفاده از سیستم‌های خودگردانِ هوشمند و فراگیری استفاده از آن را تهدیدی جدی برای جامعه بشریت معرفی نمودند. در ابتدا، خطرهایی نظیر به‌کارگیری ربات‌ها به جای انسان‌ها، قدرتمند شدنِ قدرت‌های جهانی، به‌مخاطره افتادن ارزش‌های اخلاقی و اجتماعی با جایگزینی سیستم‌های هوشمند به جای انسان‌ها و مواردی دیگر مطرح شد. اما به‌مرور، طوفان احساسات افراطی و تفریطی درباره توسعه هوش مصنوعی و پیدایش چالش‌های اخلاقی اغراق‌آمیز فرو نشست، و نگاهی واقع‌بینانه و مستدل به توانایی‌ها و محدودیت‌های هوش مصنوعی نزد بسیاری از طراحان، مهندسان و توسعه‌دهندگان این فناوری آشکار گردید. امروزه، این نگاه معتدل در قامت حوزه‌ای مطالعاتی، مسائل و چالش‌های اخلاقی مرتبط با هوش مصنوعی را از منظر اخلاق فناوری صورت‌بندی می‌کند، آن‌ها را تحلیل و تبیین می‌نماید، و سرانجام به معرفی و تدوین رهنمودهایی اخلاقی برای

توسعه هوش مصنوعی می‌پردازد (عبد خدایی و همکاران ۱۴۰۰، ۱۵-۱۶). هدف این مقاله معرفی اصول و تدوین رهنمودهای اخلاق هوش مصنوعی با مطالعه هفت سند ملی و بین‌المللی است. در این مقاله، برای دستیابی به این هدف، پنج اصل اخلاقی یکپارچه «فلورییدی» مبنای گردآوری و بررسی رهنمودهای اخلاق هوش مصنوعی قرار می‌گیرد و این رهنمودها در سه مرحله بررسی می‌شوند. در مرحله اول، رهنمودهای اخلاقی اسناد مذکور استخراج و فهرست می‌شوند. در مرحله دوم، رهنمودهای گردآوری شده، یکدست‌سازی و پالایش می‌شوند. در مرحله سوم، برای ارزیابی روایی یافته‌ها و دستاورد پژوهش، رهنمودهای اخلاقی پالایش شده ذیل اصول اخلاقی پنج‌گانه توسط گروهی کانونی از متخصصان هوش مصنوعی، فلسفه و اخلاق تکنولوژی و علم اطلاعات نقد و بررسی می‌شود تا فرایند اعتباربخشی به رهنمودهای اخلاقی برگزیده با دستیابی به اجماعی نسبی میان متخصصان تحقق یابد. در نهایت، اصول اخلاقی خیرخواهی با ۷ رهنمود، عدم آسیب‌رسانی با ۱۲ رهنمود، خودمختاری با ۷ رهنمود، عدالت با ۱۰ رهنمود، و توضیح‌پذیری با ۱۰ رهنمود دستاوردهای این پژوهش هستند.

۲. پیشینه پژوهش

سرآغاز حوزه مطالعاتی هوش مصنوعی را می‌توان به «آلن تورینگ»^۱، دانشمند علوم محاسباتی و مبدع الگوریتم رمزگشایی از ماشین رمزگذار آلمانی‌ها (به‌نام انیگما)^۲ در جنگ جهانی دوم ارجاع داد. او در سال ۱۹۵۰، در مقاله‌ای که با عنوان «دستگاه محاسبات و هوش» در مجله Mind منتشر شد، برای نخستین بار به مفهوم هوش ماشین اشاره کرد و ادعا کرد که ماشین‌ها هم می‌توانند مانند انسان فکر کنند. در ۱۹۵۶، در کنفرانس دارتموث اصطلاح هوش مصنوعی توسط «جان مک‌کارتی»^۳ دانشمند علوم محاسباتی و علوم شناختی به کار برده شد. در این کنفرانس «هربرت سایمون»^۴ و «آلن نیوول»^۵ برنامه‌ای ارائه کردند که با استفاده از هوش مصنوعی و بهره‌گیری از زیربناهای منطق ریاضی می‌توانست قضیه‌های اصولی ریاضیات را اثبات کند. از ۱۹۷۰ تا ۱۹۸۰، محققان، سیستم‌های هوشمند بیشتری

1. Alan Turing

2. Enigma

3. John McCarthy

4. Herbert Simon

5. Allen Newell

با کاربرد در زیست‌شناسی، پزشکی، مهندسی، و صنایع نظامی طراحی کردند. در سال ۲۰۰۰، یادگیری ماشین آماری که از دهه ۸۰ میلادی آغاز شده بود، کاربردهای گسترده‌ای در خدمات نرم‌افزاری و دستگاه‌های موبایل پیدا کرد. در سال ۲۰۰۹، دانشگاه استنفورد «انجمن پیشبرد هوش مصنوعی»^۱ را برای مطالعه آینده هوش مصنوعی پایه‌گذاری کرد. امروزه، در ادامه تحولات مستمر هوش مصنوعی، ربات چت هوش مصنوعی با عنوان ChatGPT گسترش و توسعه یافته است. این ربات توسط شرکت OpenAI توسعه یافته است. این شرکت در سال ۲۰۱۵ تأسیس شد و از سوی سرمایه‌گذاران بزرگی همچون مایکروسافت پشتیبانی می‌شود. ربات ChatGPT یک سرویس مبتنی بر متن است که پاسخ‌هایی شبیه به انسان برای درخواست‌ها و پرسش‌های کاربر فراهم می‌کند. این ربات به هر سؤال پاسخ داده و به معنای واقعی کلمه همه‌فن‌حریف است (بهاری ۱۳۹۸، ۲۱-۲۴). اگرچه این تحولات هر روزه، در حوزه هوش مصنوعی و امکانات و ابزارهای مبتنی بر آن همچنان ادامه دارد، اما به موازات این توسعه، حساسیت‌ها و ملاحظات اخلاقی گوناگونی شکل می‌گیرند. به بیان دیگر، هر اندازه که ماشین‌های هوشمند بتوانند به‌طور فعال با انسان‌ها در تعامل باشند، به همان اندازه می‌توانند موقعیت‌های اخلاقی پیچیده و گاهی نوظهور ایجاد کرده و نقش مؤثری ایفا کنند. امروزه، ویژگی‌ها و کارکردهای انسانی به ربات‌های هوشمند انتقال داده می‌شوند و به دنبال آن در موقعیت‌های اخلاقی، نقش مشابه یک انسان برای آن‌ها در نظر گرفته می‌شود. به عنوان مثال، یک پرستار در قبال بیمار خود مسئولیت‌های اخلاقی مشخصی دارد. حال اگر همین پرستار یک ربات هوشمند باشد، آیا همان مسئولیت‌های اخلاقی بر عهده این ربات هوشمند هم خواهد بود؟

در پاسخ به این پرسش می‌توان سه رویکرد را اتخاذ کرد: رویکرد اول) این گونه موقعیت‌ها برای ماشین هوشمند، اخلاقی محسوب نمی‌شود و اگر به این گونه شرایط صفت اخلاقی نسبت دهیم، از اساس دچار خطا شده‌ایم. رویکرد دوم) چنین موقعیتی شبه‌اخلاقی^۲ است؛ به این معنا که بخشی از آن اخلاقی بوده، اما عامل‌های هوشمند ماشینی در این شرایط فاقد ویژگی‌های به‌خصوصی هستند که یک عامل اخلاقی تام^۳ باید واجد آن باشد. بنابراین، هر چند می‌توان آن‌ها را به نوعی عامل اخلاقی در نظر گرفت، ولی عامل

1. Association for the Advancement of Artificial Intelligence (AAAI)

2. pseudo-moral

3. full moral agent

اخلاقی تام محسوب نمی‌شوند. رویکرد سوم) چنین موقعیتی اخلاقی محسوب می‌شود و باید به‌طور جدی و به‌عنوان یک مسئله اخلاقی به آن نگریست.

پیرامون این رویکردهای سه‌گانه، مجموعه آثار و دیدگاه‌های مختلفی در اخلاق هوش مصنوعی گردآوری و اتخاذ شده است. در سال‌های اخیر رشد و انباشت منابع تخصصی در این حوزه از چنان شتابی برخوردار شده است که مروری بر کتاب‌ها و مجموعه مقالات گردآوری شده صرفاً در چند سال اخیر، گواهِ اهمیت این حوزه مطالعاتی و پرداختن به مسائل و چالش‌های اخلاقی مرتبط است. در ادامه به معرفی برخی از این آثار پرداخته می‌شود.

«پائولا بودینگتون» در کتاب خود «به‌سوی کدنام‌های اخلاقی برای هوش مصنوعی»، با توجه به سرعت روزافزون شکل‌گیری مسائل اخلاقی فعلی و آتی در توسعه هوش مصنوعی، به بررسی چگونگی تهیه و تدوین کدها یا مقررات اخلاقی واقعی و قابل اجرا در این زمینه می‌پردازد. او در این اثر ملاحظات اخلاقی کلیدی را به‌اختصار بیان می‌کند، همه جوانب استدلال‌ها را آشکار می‌سازد، و به چگونگی تدوین کدهای اخلاقی می‌پردازد (Boddington 2017). این کتاب منبع مفیدی برای کسانی است که قصد دارند چالش‌های اخلاقی پژوهش‌های هوش مصنوعی را به روش‌هایی معنادار و عملی بررسی و مطالعه کنند.

«مارک کو کلب‌رگ» در کتاب «اخلاق هوش مصنوعی»، به‌عنوان یک فیلسوف فناوری، فراتر از سناریوهای ژورنالیستی و عامه‌پسند، روایت‌های قابل تأملی از هوش مصنوعی را توصیف می‌کند؛ از هیولای فرانکنشتاین^۱ تا فرانس‌گرایی^۲ و تکنیکی فناورانه^۳. وی بحث‌های فلسفی متعدد مرتبطی را بررسی می‌کند و پرسش‌هایی درباره تفاوت‌های اساسی بین انسان و ماشین و بحث‌هایی درباره وضعیت اخلاقی هوش مصنوعی پیش می‌آورد. او فناوری‌های هوش مصنوعی را توضیح می‌دهد و رویکردهای مختلف آن را توصیف کرده و بر حوزه‌های یادگیری ماشین و علم داده تمرکز می‌کند. در این اثر مروری بر مسائل مهم اخلاقی، از جمله نگرانی‌های مربوط به حریم خصوصی، مسئولیت و تفویض اختیار

1. Frankenstein's monster
2. transhumanism

۳. technological singularity اصطلاح سینگولاریتی تکنولوژیکی یا تکنیکی فناورانه به نقطه‌ای فرضی در آینده اشاره دارد که در آن رشد هوش مصنوعی از توان شناختی انسان پیشی می‌گیرد و سپس پیشرفت فناوری به‌صورت نمایی، غیرقابل پیش‌بینی و خارج از کنترل انسانی ادامه می‌یابد.

تصمیم‌گیری، شفافیت و سوگیری در تمامی مراحل فرایندهای علم داده ارائه می‌شود. در این اثر همچنین آینده کار در اقتصاد هوش مصنوعی بررسی می‌گردد. در نهایت، نویسنده طیفی از پیشنهاد‌های سیاستی را تجزیه و تحلیل می‌کند و چالش‌های پیش‌روی سیاست‌گذاران در این حوزه را بحث و بررسی می‌کند. در پایان، وی به دفاع از رویه‌های اخلاقی می‌پردازد که ارزش‌های تعبیه‌شده در طراحی‌های فناورانه را به ارزش‌های دموکراتیک درون رویه‌ها تبدیل می‌کند و استدلال می‌کند که از این راه به چشم‌اندازی از زندگی خوب و جامعه خوب دست می‌یابیم (Coeckelbergh 2020).

در اثری با نام «مقدمه‌ای بر اخلاق در رباتیک و هوش مصنوعی»، مخاطبان با مبانی هوش مصنوعی و اخلاق آشنا می‌شوند. در این اثر درباره مسائل اعتماد، مسئولیت‌پذیری، حریم خصوصی و مخاطرات در رابطه با هوش مصنوعی بحث می‌شود. محتوای این کتاب به گونه‌ای ارائه شده است که مخاطبان نیاز گسترده‌ای به تخصص فنی، حقوقی یا فلسفی نیاز ندارند. نویسندگان از مثال‌های متعددی برای بیان مسائل و چالش‌های اخلاقی استفاده می‌کنند و کتاب را با بحث در زمینه‌های کاربردی هوش مصنوعی و رباتیک، به ویژه وسایل نقلیه خودران، سیستم‌های تسلیحات خودکار و الگوریتم‌های سوگیرانه به پایان می‌رسانند (Bartneck & et al. 2021).

در اثر دیگری از «پائولا بودینگتون» که در قالب کتابی درسی با عنوان «اخلاق هوش مصنوعی» تدوین شده است، مخاطبان با دغدغه‌های اخلاقی مهمی در توسعه و استفاده از هوش مصنوعی آشنا می‌شوند. او با ارائه اطلاعاتی روشن و قابل فهم درباره مفاهیم و موضوعات اصلی در اخلاق هوش مصنوعی، به بررسی این مسئله می‌پردازد که چالش‌های اخلاقی فعلی مرتبط با توسعه هوش مصنوعی چگونه ما را وادار می‌کنند که به پرسش‌های اساسی و قدیمی درباره زندگی انسان، ارزش و معنا بپردازیم. افزون بر این، این کتاب نشان می‌دهد که چگونه مسائل بنیادی و نظری در حوزه اخلاق با مناقشات عینی اخلاقی در حوزه هوش مصنوعی مرتبط هستند. در ادامه، پرسش‌های مربوط به اخلاق هوش مصنوعی در زمینه موضوعات مرتبط با فناوری، مقررات، جامعه، مذهب و فرهنگ بررسی می‌شوند تا درک روشنی از حوزه اخلاق هوش مصنوعی به خوانندگان عرضه شود (Boddington 2022).

در کتاب «اخلاق هوش مصنوعی: مطالعات موردی و گزینه‌هایی برای رسیدگی به چالش‌های اخلاقی»، نویسندگان، این مجموعه را با هدف دسترسی مطلوب به مطالعات

موردی در حوزه اخلاق هوش مصنوعی گردآوری کرده‌اند. از این‌رو، اولین کتابی است که مطالعات موردی زندگی واقعی را همراه با تفسیرها و راهبردهایی برای غلبه بر چالش‌های اخلاقی ارائه می‌کند. مطالعات موردی یکی از بهترین روش‌ها برای فهم چالش‌های اخلاقی هوش مصنوعی و دستیابی به بینشی مناسب درباره پیچیدگی‌های مختلف و دیدگاه‌های بهره‌داران این چالش‌هاست. با توجه به حضور همه‌جانبه اخلاق هوش مصنوعی در بحث‌های دانشگاهی، سیاست‌گذاری، و رسانه‌ای، این کتاب برای طیف وسیعی از مخاطبان، از محققان رشته‌های مختلف (مانند علم هوش مصنوعی، اخلاق، سیاست، فلسفه، اقتصاد) گرفته تا مدیران و سیاست‌گذاران مناسب خواهد بود (Stahl, Schroeder & Rodrigues 2022).

در کتاب «اخلاق هوش مصنوعی برای اهداف توسعه پایدار»، نویسندگان شامل متخصصانی هستند که با دیدگاه‌هایی از حوزه‌های مختلف به گردآوری این مجموعه پرداخته‌اند. در فصل‌های مختلف این مجموعه، مشکلات حاد حاکمیتی مربوط به استفاده از هوش مصنوعی برای اهداف توسعه پایدار معرفی و مطالعاتی موردی درباره اینکه چگونه هوش مصنوعی در حال پیشرفت می‌تواند به اهداف توسعه پایدار دست یابد، بحث و بررسی می‌شود. محققانی که در رابطه با هوش مصنوعی برای اهداف توسعه پایدار و حاکمیت اخلاقی هوش مصنوعی پژوهش می‌کنند، مخاطبان اصلی این اثر هستند (Mazzi & Floridi 2023).

آثار معتبر و روزآمد فوق تنها نمونه بسیار اندکی از اسناد، دستورالعمل‌ها، مقالات و دیگر منابع تخصصی در حوزه اخلاق هوش مصنوعی هستند که ضرورت شناسایی و اهمیت تدوین اصول و رهنمودهای اخلاقی برای مواجهه با چالش‌های اخلاقی در توسعه هوش مصنوعی را آشکار می‌کنند.

۳. روش پژوهش

همان‌گونه که اشاره شد، هدف این مقاله معرفی اصول و تدوین رهنمودهای اخلاقی هوش مصنوعی با مطالعه هفت سند ملی و بین‌المللی است. در این مقاله برای دستیابی به این هدف، پنج اصل اخلاقی یکپارچه «لوچیانو فلورییدی»، فیلسوف معاصر فلسفه و اخلاق

اطلاعات، مبنای گردآوری و بررسی رهنمودهای اخلاق هوش مصنوعی قرار می‌گیرد و در سه مرحله بررسی می‌شود. در مرحله اول، رهنمودهای اخلاقی اسناد مذکور استخراج و فهرست می‌شوند. در مرحله دوم، رهنمودهای گردآوری شده یکدست‌سازی، موارد تکراری حذف، و رهنمودهای جزئی‌تر به نفع رهنمودهای کلی‌تر کنار گذاشته می‌شوند. در مرحله سوم، برای ارزیابی روایی یافته‌ها و دستاورد پژوهش، رهنمودهای اخلاقی پالایش شده ذیل اصول اخلاقی پنج‌گانه توسط گروهی کانونی از متخصصان هوش مصنوعی، فلسفه و اخلاق فناوری و علم اطلاعات نقد و بررسی می‌شود. این بررسی‌ها تا زمان دستیابی به اجماعی نسبی میان متخصصان گروه کانونی ادامه داشته و فرایند اعتباربخشی به یافته‌های پژوهش تحقق می‌یابد.

۴. معرفی اصول اخلاقی یکپارچه برای اخلاق هوش مصنوعی

حوزه مطالعاتی اخلاق هوش مصنوعی اغلب بر انواع مختلفی از چالش‌های اخلاقی و اجتماعی ناشی از توسعه و استقرار سیستم‌های هوش مصنوعی متمرکز است و دربردارنده مجموعه‌ای از ملاحظات و پاسخ‌هایی نقادانه به چالش‌هاست. البته طیف وسیعی از این چالش‌ها اغراق‌آمیز یا دست‌کم مسائلی اخلاقی هستند که نوظهور نیستند و پیش از پیدایش مصنوعات مرتبط با هوش مصنوعی وجود داشته‌اند و صرفاً بازآفرینی یا تشدید شده‌اند. اما برخی از آن‌ها در حقیقت، منبعت از توسعه و استقرار سیستم‌های هوش مصنوعی هستند. تفکیک میان این دو دسته چالش‌های ظاهری و حقیقی نیازمند تحلیل و بررسی نقادانه و موشکافانه است.

مرامنامه‌ها یا اسناد اخلاقی برای توسعه و استقرار هوش مصنوعی به مجموعه‌ای از اصول اخلاقی و رهنمودهای اخلاقی گفته می‌شود که در چارچوب اخلاق کاربردی برای مواجهه با چالش‌های اخلاقی و اجتماعی ناشی از طراحی، توسعه و استقرار سیستم‌های هوش مصنوعی گردآوری، تبیین و استنتاج می‌شود. این مرامنامه‌ها یا اسناد اخلاقی به‌طور معمول، با اجماعی سازمانی یا نهادی ضمانت اجرایی می‌یابند. بنابراین، آن‌ها را می‌توان در قالب اسناد ملی یا بین‌المللی در رابطه با اخلاق هوش مصنوعی به تصویب رسانده و اجرایی کرد. این اسناد به‌گفته دیگر، مجموعه‌ای از اصول و رهنمودهای اخلاقی مرتبط با اخلاق هوش مصنوعی است که با همکاری میان مهندسان هوش مصنوعی، محققان اخلاق، روان‌شناسان، جامعه‌شناسان، حقوق‌دانان، سیاست‌گذاران و برخی دیگر از متخصصان

مرتبط با این حوزه برای مواجهه اخلاقی در توسعه و استقرار سیستم‌های هوشمند در سطح ملی، منطقه‌ای و جهانی گردآوری و ارائه می‌شود (Gordon & Wrenn 20).

از این رو، در چند دهه اخیر با رشد روزافزون سیستم‌های هوش مصنوعی، نهادها و سازمان‌های با نفوذ ملی یا بین‌المللی (از جمله دولت‌ها، یونسکو، اتحادیه اروپا، شرکت‌ها و انجمن‌های تخصصی) وظیفه تهیه پیش‌نویس منشورها و رهنمودهای اخلاقی با هدف تحقق سیستم‌های هوش مصنوعی مفید برای جامعه را به گروه‌های متخصص در حوزه‌های مطالعاتی فنی، حقوقی، اجتماعی، اخلاقی مرتبط با هوش مصنوعی واگذار کرده‌اند. امروزه، چنین اسنادی به‌خاطر تعدد و تکثر سازمان‌ها و نهادهای ملی و بین‌المللی مرتبط با توسعه و استقرار اخلاقی سیستم‌های هوش مصنوعی به حدی گسترش یافته‌اند که در آن پیگیری آخرین اصول و رهنمودهای منتشرشده بسیار دشوار است. افزون بر این، سازمان‌های مرتبط با اخلاق هوش مصنوعی بودجه قابل توجهی از منابع مختلف دولتی و خصوصی دریافت می‌کنند و چندین مرکز تحقیقاتی برای این منظور تأسیس شده است. این تحولات به‌طور عمده، پاسخ‌های مثبتی دریافت کرده است. در همین حال، به‌خاطر تکثر اصول و رهنمودهای اخلاق هوش مصنوعی، نگرانی‌هایی نسبت به شست‌وشوی اخلاقی^۱ فعالیت‌های پژوهشی و تکنولوژیکی ناصواب در شرکت‌های تخصصی وجود دارد^۲. اما یکی از مهم‌ترین پرسش‌های قابل طرح این است که آیا می‌توان میان اصول اخلاقی و به‌دنبال آن، میان رهنمودهای متفاوت در کدهای اخلاق هوش مصنوعی اتفاق نظر ایجاد کرد یا خیر (Gordon & Wrenn 2020).

هدف این مقاله معرفی اصول یکپارچه اخلاقی و تدوین رهنمودهای اخلاقی هوش مصنوعی با مطالعه هفت سند ملی و بین‌المللی در چارچوب این اصول اخلاقی است. برای دستیابی به این هدف، پنج اصل اخلاقی «لوچیانو فلوریدی»، فیلسوف معاصر فلسفه و اخلاق اطلاعات، مبنای گردآوری و بررسی رهنمودهای اخلاقی هوش مصنوعی قرار می‌گیرد. اهمیت معرفی اصول یکپارچه اخلاقی بدین‌خاطر است که به‌منظور گردآوری، تدوین، و پیشنهاد رهنمودهای اخلاقی هوش مصنوعی لازم است اصول بنیادینی که رهنمودهای

1. ethics washing

۲. اصطلاح شست‌وشوی اخلاق، به موقعیتی اشاره دارد که در آن، سازمان یا شرکت بیش از آنکه دغدغه ارتقای اخلاقی خدمات یا محصولات خود را داشته باشد، به‌دنبال ضرب‌مهر تأیید اخلاقی بر این فعالیت‌هاست، حال آنکه با نگاهی انتقادی‌تر و عمیق‌تر، فعالیت‌های سازمانی مذکور صرفاً از ظاهری اخلاقی برخوردار است

اخلاقی با استناد به آن‌ها تهیه و تنظیم می‌شوند، متمایز و معرفی شوند. در اینجا، با محور قرار دادن نتایج پژوهشی مقاله‌ای از «فلوریدی و کولز» با عنوان «چارچوبی یکپارچه از پنج اصل برای هوش مصنوعی در جامعه» (Floridi & Cowl 2019)، در نظر داریم اصول اخلاقی یکپارچه و بنیادی برگرفته از این پژوهش را مبنای گردآوری و بررسی رهنمودهای اخلاقی هوش مصنوعی قرار دهیم.

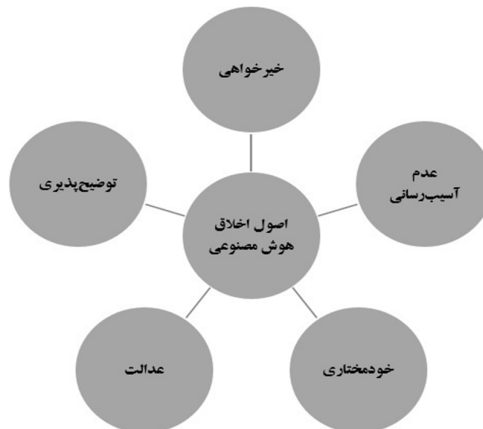
برای پیشنهاد رهنمودهای اخلاقی، سازمان‌ها و نهادهای ملی و بین‌المللی متعددی اقدام به گردآوری و تدوین اصول اخلاقی هوش مصنوعی کرده‌اند. شوربختانه، حجم بالای اصول پیشنهادی انبوهی از این اصول و تهدیدی برای سردرگمی به بار آورده است. این وضعیت دو مسئله بالقوه را به وجود می‌آورد: اگر مجموعه‌های مختلفی از اصول اخلاقی برای هوش مصنوعی، مشابه یکدیگر باشند، این مسئله منجر به تکرار و افزونگی غیرضروری می‌شود؛ اما اگر آن‌ها به‌طور قابل توجهی متفاوت از همدیگر باشند، منجر به سردرگمی و ابهام می‌شود. در سال ۲۰۱۷، پس از انتشار اصول هوش مصنوعی «آسیلومار»^۱ و اعلامیه مونترال برای توسعه مسئولانه هوش مصنوعی، بسیاری از سازمان‌ها و نهادهای ملی و بین‌المللی در جهت عمل به مسئولیت اجتماعی و البته تثبیت اقتدار نهادی خود، اقدامات وسیعی برای تدوین اصول اخلاقی هوش مصنوعی به عمل آورده‌اند. این اسناد اخلاقی عبارت‌اند از: (۱) اصول هوش مصنوعی «آسیلومار» (Future of Life Institute 2017)؛ (۲) بیانیه مونترال برای توسعه مسئولانه هوش مصنوعی (Université de Montréal 2018)؛ (۳) طراحی همسو: چشم‌اندازی برای اولویت دادن به رفاه انسان همراه با سیستم‌های خودگردان و هوشمند «مؤسسه مهندسان برق و الکترونیک» (IEEE 2017)؛ (۴) اصول اخلاقی بیانیه هوش مصنوعی، رباتیک و سیستم‌های خودگردان؛ گروه اروپایی کمیسیون اروپا در زمینه اخلاق در علم و تکنولوژی‌های نوظهور (High-Level Expert Group on Artificial Intelligence)؛ (۵) پنج اصل کلی برای کدهای هوش مصنوعی در گزارش مجلس بریتانیا (House of Lords Select Committee on Artificial Intelligence 2018)؛ (۶) بیانیه اصول اخلاقی مؤسسه مشارکت در هوش مصنوعی (Partnership on AI 2016)؛ و (۷) توصیه‌نامه اخلاق هوش مصنوعی کنفرانس عمومی یونسکو (UNESCO 2021).

امروزه با حجم روزافزون اسناد اخلاق هوش مصنوعی، خطر تکرار و همپوشانی

1. Asilomar

غیر ضروری در صورت مشابه بودن مجموعه‌های مختلف اصول، یا سردرگمی و ابهام در صورت متفاوت بودن آن‌ها پدیدار شده است. از نظر «فلورییدی و کولز» زمان آن فرا رسیده است که یک تحلیل مقایسه‌ای معطوف به این اسناد انجام شود. از موارد قابل بررسی در تحلیل آنان، ارزیابی این موضوع است که آیا این اصول اخلاقی متعدد همگرا هستند یا واگرا؟ و اگر همگرا باشند، می‌توان یک چارچوب واحد و یکپارچه برای این اصول تنظیم و ارائه کرد. در پژوهش آنان، این اسناد بر اساس چهار معیار اساسی انتخاب شده‌اند که عبارت‌اند از: الف) متأخر بودن؛ به این معنا که اسناد مذکور از سال ۲۰۱۷ به بعد منتشر شده‌اند؛ ب) به‌طور مستقیم مرتبط با هوش مصنوعی و تأثیر آن بر کلیت جامعه هستند؛ ج) بسیار معتبر هستند و از اقتدار نهادی ملی و بین‌المللی برخوردارند؛ و د) به دلیل دارا بودن ویژگی‌های الف تا ج تأثیرگذار هستند. بر این اساس، «فلورییدی و کولز» اسناد معرفی شده ۱ تا ۶ را از منظر همگرایی یا واگرایی اصول اخلاقی مندرج در آن‌ها، و در صورت همگرایی آن‌ها، ارائه چارچوب یکپارچه‌ای از اصول اخلاقی که غیرقابل فروکاست به یکدیگر هستند، بررسی تحلیلی و نقادانه کردند (Floridi & Cowls 2019).

به‌طور کلی، ۶ سند نخست، نزدیک به ۴۷ اصل را ارائه می‌کنند. اگرچه، تفاوت‌ها در این اصول به‌طور عمده زبانی هستند و درجه‌ای از هماهنگی و همپوشانی قابل توجهی بین این اصول متعدد و متکثر وجود دارد. در نهایت، نتیجه بررسی‌های آنان به معرفی ۵ اصل اخلاقی یکپارچه، مجزا و غیرقابل فروکاست به یکدیگر منجر می‌شود. در شکل ۱، اصول اخلاقی مذکور نشان داده شده است.



شکل ۱. پنج اصل اخلاقی مجزا و یکپارچه اخلاق هوش مصنوعی از دیدگاه «فلورییدی و کولز»

همان‌گونه که اشاره شد، برای اجتناب از تکرار و افزونگی غیرضروری در تکرار و تنوع اصول اخلاقی از سویی، و رفع سردرگمی و ابهام ناخواسته در تعاریف همپوشانی‌کننده در اصول اخلاقی از سوی دیگر، ما را به پنج اصل اخلاقی برگرفته از تحقیقات «فلوریدی و کولز» رهنمون ساخت. اکنون ذیل این اصول پنج‌گانه اخلاقی، رهنمودهای هفت سند اخلاق هوش مصنوعی را بررسی و به ترتیب در سه مرحله به گردآوری، پالایش، و اعتباربخشی آن‌ها می‌پردازیم.

برای تحقق مراحل سه‌گانه فوق ابتدا لازم است این اصول اخلاقی پنج‌گانه، فهرست شده، قلمرو مفهومی آن‌ها مشخص گردد، و برای آن‌ها تعریفی صریح ارائه شود. از این طریق رهنمودهای اخلاقی را در سه مرحله این پژوهش به نحوی دقیق‌تر مورد بررسی قرار می‌دهیم. در جدول ۱، هر اصل اخلاقی تعریف شده و قلمرو مفهومی آن‌ها برای پرهیز از فروکاست به اصل اخلاقی دیگری مشخص شده است.

جدول ۱. تعریف پنج اصل اخلاقی و قلمرو مفهومی آن برای بررسی رهنمودهای اخلاق هوش مصنوعی

اصل اخلاقی	قلمرو مفهومی	تعریف
خیرخواهی ^۱	ارتقای رفاه، حفظ کرامت و پایداری کره زمین	هوش مصنوعی برای خیر اجتماعی ^۲ ، در خدمت منافع عمومی برای بشریت است و در نهایت، موجب ارتقای رفاه و حفظ کرامت انسانی و پایداری محیط زیست می‌شود.
عدم آسیب‌رسانی ^۳	حریم خصوصی، امنیت و احتیاط	هوش مصنوعی برای خیر اجتماعی، موجب نقض حریم خصوصی و امنیت انسان‌ها نمی‌شود و از توانمندی آن علیه بشریت استفاده نمی‌گردد.
خودمختاری ^۴	قدرت تصمیم‌گیری برای تصمیم‌گیری	هوش مصنوعی برای خیر اجتماعی، حق انتخاب و قدرت تصمیم‌گیری انسان‌ها را محدود یا دچار اختلال نمی‌کند، بلکه موجب ارتقای آن‌ها می‌شود.
عدالت ^۵	گسترش رفاه، حفظ همبستگی، اجتناب از بی‌عدالتی	هوش مصنوعی برای خیر اجتماعی، در جهت گسترش رفاه، حفظ همبستگی، اجتناب از تبعیض‌ها، فرصت دسترسی برابر به رفاه و منافع مشترک برای انسان‌ها را فراهم می‌آورد.
توضیح‌پذیری ^۷	امکان‌پذیری دیگر اصول از طریق درک‌پذیری و پاسخگویی	هوش مصنوعی برای خیر اجتماعی، دارای فرایندهایی شفاف است؛ به‌نحوی که قابلیت‌ها و هدف این سیستم‌ها را به‌طور شفاف به یکدیگر مرتبط می‌کند و تصمیمات این سیستم‌ها تا حد امکان برای افرادی که به‌صورت مستقیم یا غیرمستقیم تحت تأثیر آن‌ها قرار می‌گیرند، قابل توضیح است.

1. beneficence

2. AI for social good: AI4SG

3. nonmaleficence

4. autonomy

5. justice

6. solidarity

7. explicability

۵. تجزیه و تحلیل هفت سند اخلاق هوش مصنوعی

تاکنون به معرفی ۵ اصل اخلاقی یکپارچه پرداخته شده است. اکنون در نظر داریم با فهرست کردن هفت سند ملی و بین‌المللی در حوزه اخلاق هوش مصنوعی با معیارهای اعتبارسنجی که در قسمت ۴ به آن‌ها اشاره شد، ذیل اصول اخلاقی مذکور، رهنمودهای اخلاقی این اسناد را استخراج و ارزیابی کنیم. این رهنمودها در سه مرحله گردآوری، پالایش، و اعتباربخشی بررسی می‌شوند. ابتدا لازم است یادآوری شود که ۶ سند نخست از اسناد اخلاق هوش مصنوعی در قسمت ۴، همان اسنادی هستند که «فلوریدی» در تحقیق خود مبنای استخراج ۵ اصل اخلاقی قرار داده بود و به رهنمودهای اخلاقی آن‌ها، در صورت وجود، پرداخته بود. سند هفتم به این مجموعه اسناد در این مقاله، به دلیل اهمیت، روزآمدی، اعتبار و جهان‌شمولی آن افزوده شده است. این سند در واقع، توصیه‌نامه‌ای مبسوط در خدمت اخلاق هوش مصنوعی است که در جلسه چهل و یکم کنفرانس عمومی یونسکو پس از بحث و بررسی نمایندگان کشورهای عضو، در نوامبر ۲۰۲۱ به تصویب رسید و به‌عنوان سندی مشورتی در اختیار کشورها قرار گرفت.

در مرحله اول ارزیابی، رهنمودهای اخلاقی اسناد مذکور استخراج و فهرست می‌شوند. در مرحله دوم، رهنمودهای گردآوری‌شده یکدست‌سازی، موارد تکراری حذف، و رهنمودهای جزئی‌تر به‌نفع رهنمودهای کلی‌تر کنار گذاشته می‌شوند. در مرحله سوم، برای ارزیابی روایی یافته‌ها و دستاورد پژوهش، رهنمودهای اخلاقی پالایش‌شده ذیل اصول اخلاقی پنج‌گانه توسط گروهی کانونی از متخصصان هوش مصنوعی، فلسفه و اخلاق فناوری و علم اطلاعات نقد و بررسی می‌شود. این بررسی‌ها تا زمان دستیابی به اجماعی نسبی میان متخصصان گروه کانونی ادامه یافته و فرایند اعتباربخشی به یافته‌های پژوهش تحقق می‌یابد.

۵-۱. مرحله اول: استخراج رهنمودهای اخلاقی از هفت سند اخلاق هوش مصنوعی

رهنمودهای اخلاقی در بررسی هفت سند اخلاق هوش مصنوعی از طریق جست‌وجوی کلمات کلیدی یکسان و مشابه برای هر یک از اصول اخلاقی پنج‌گانه استخراج و گردآوری شدند. به‌عنوان مثال، برای اصل خیرخواهی، افزون بر واژه کلیدی beneficence و مشتقات و مترادف‌های آن، و کلماتی مانند wellbeing, dignity و sustainability نیز در تمامی این اسناد جست‌وجو شدند و رهنمودهایی که برخوردار از این واژه‌های کلیدی

بودند، استخراج و گردآوری شدند. در مجموع، ۱۳۰ رهنمود ذیل ۵ اصل اخلاقی از ۷ سند اخلاق هوش مصنوعی استخراج شدند. ۱۵ رهنمود برای اصل خیرخواهی، ۴۹ رهنمود برای اصل عدم آسیب‌رسانی، ۲۵ رهنمود برای اصل خودمختاری، ۲۸ رهنمود برای اصل عدالت، و ۲۸ رهنمود برای اصل توضیح‌پذیری از این اسناد گردآوری شدند.

۲-۵. مرحله دوم: یکدست‌سازی و پالایش رهنمودهای گردآوری‌شده

در مرحله دوم، رهنمودهای ویرایش از نظر اصطلاحات یکدست‌سازی شدند. همچنین رهنمودهای تکراری حذف و رهنمودهایی که از جامعیت کمتری نسبت به رهنمودهای دیگر برخوردار بودند، حذف شدند. به این ترتیب، در مجموع، ۶۷ رهنمود در قالب ۸ رهنمود برای اصل خیرخواهی، ۱۶ رهنمود برای اصل عدم آسیب‌رسانی، ۱۳ رهنمود برای اصل خودمختاری، ۱۴ رهنمود برای اصل عدالت، و ۱۶ رهنمود برای اصل توضیح‌پذیری حفظ شدند. در مرحله بعد، رهنمودهای موجود برای ارزیابی و اعتباریابی در گروه کانونی از متخصصان مورد بحث و بررسی قرار گرفتند.

۶. یافته‌های پژوهش: مرحله سوم و برگزاری گروه کانونی اخلاق هوش مصنوعی

در این مرحله رهنمودهای اخلاقی به‌دست‌آمده از مرحله دوم در گروه کانونی بحث و بررسی گردید. این گروه شامل ۶ نفر متخصص در حوزه هوش مصنوعی و اخلاق حرفه‌ای (دو متخصص علم اطلاعات و دانش‌شناسی، یک متخصص فلسفه علم و فناوری، یک متخصص مهندسی فناوری اطلاعات، یک متخصص مهندسی صنایع، و یک متخصص سیاست‌گذاری بر پایه مدل) برگزار شد. در این نشست، ۶۷ رهنمود اخلاقی مرحله دوم در چارچوب پنج اصل اخلاقی بار دیگر اولویت‌بخشی، اصلاح و صورت‌بندی شدند. در جدول ۲، اصول اخلاقی و رهنمودهای پیشنهادی در گروه کانونی این پژوهش، جمع‌بندی و فهرست می‌شود.

جدول ۲. اصول و رهنمودهای پیشنهادی در اخلاق هوش مصنوعی

ردیف	اصل اخلاقی	رهنمودهای اخلاقی مرتبط
۱	خیرخواهی (با ۷ رهنمود)	<ul style="list-style-type: none"> ◇ سیستم‌های هوش مصنوعی باید به‌گونه‌ای توسعه و استقرار یابند که رفاه انسانی را به‌عنوان پیامد طراحی این سیستم‌ها تضمین کند. ◇ سیستم‌های هوش مصنوعی باید به‌گونه‌ای توسعه و استقرار یابند

ردیف اصل اخلاقی رهنمودهای اخلاقی مرتبط

- ◆ که با خیر اجتماعی و پایداری محیط زیست در تعارض قرار نگیرند.
- ◆ سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که امکان رشد و بهبود برای رفاه تمام موجودات صاحب شعور^۱ فراهم شود.
- ◆ سیستم‌های هوش مصنوعی باید در خدمت منافع و توانمندسازی بیشترین افراد ممکن باشد.
- ◆ سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که با آرمان‌های کرامت انسانی، حقوق بشر، آزادی‌ها، خیر عمومی و تنوع فرهنگی سازگار باشند.
- ◆ فرایندهای توسعه و استقرار سیستم‌های هوش مصنوعی باید ذیل چارچوب‌های حکمرانی (شامل استانداردها و نهادهای نظارتی) ارزیابی و ردیابی شوند تا حقوق، آزادی‌ها، کرامت، و حریم خصوصی انسان‌ها را نقض نکنند.
- ◆ ارزیابی تأثیرات اجتماعی، اقتصادی، زیستی سیستم‌های هوش مصنوعی باید با آگاهی کامل از پیامدهای این تکنولوژی برای توسعه پایدار به گونه‌ای پیوسته و در حال تکامل انجام شود.
- ◆ حریم خصوصی انسان‌ها در طول چرخه عمر سیستم‌های هوش مصنوعی باید به‌عنوان حقی ضروری برای حفاظت از کرامت انسانی، خودمختاری، و عاملیت انسانی رعایت، محافظت و ترویج شود.
- ◆ توسعه سیستم‌های هوش مصنوعی باید در جهت ایجاد منافع اجتماعی و اقتصادی افراد از طریق کاهش نابرابری‌ها و آسیب‌پذیری‌های اجتماعی باشد.
- ◆ سیستم‌های هوش مصنوعی باید در طول چرخه عمر خود ایمن و قابل اعتماد باشند و این ایمنی قابل تأیید باشد.
- ◆ در تعامل با سیستم‌های هوش مصنوعی، افراد باید حق دسترسی، مدیریت، و کنترل بر داده‌های تولیدشده خود را داشته باشند.
- ◆ سیستم‌های هوش مصنوعی باید در جمع‌آوری و بایگانی داده‌ها، حق محرمانگی داده‌ها و ناشناس بودن پروفایل‌های شخصی را تضمین کنند.
- ◆ حریم خصوصی افراد باید از رخنه سیستم‌های هوش مصنوعی و سیستم‌های جمع‌آوری و بایگانی داده‌ها محافظت شود.
- ◆ سیستم‌های هوش مصنوعی باید حق خلوت و امکان جدایی خودخواسته افراد از ارتباط و تعامل با آن‌ها را فراهم کنند.
- ◆ در توسعه سیستم‌های هوش مصنوعی باید با حفظ احتیاط، پیامدهای منفی آن‌ها را به‌نحوی مسئولانه و تا حد امکان پیش‌بینی و اقدامات پیش‌گیرانه را اتخاذ کرد.
- ◆ نسبت به خطرات بالقوه سوء استفاده از سیستم‌های هوش مصنوعی باید با ارائه آموزش‌های اخلاقی و امنیتی، حساسیت افراد جامعه را ارتقا داد.

۲ عدم آسیب‌رسانی (با ۱۲ رهنمود)

1. sentient beings

ردیف	اصل اخلاقی	رهنمودهای اخلاقی مرتبط
		<ul style="list-style-type: none"> ◇ سیستم‌های هوش مصنوعی باید برای اطمینان از حفظ امنیت جسمی، ذهنی، و محیطی انسان‌ها، پیش از عرضه به دقت آزمایش شوند. ◇ آسیب‌های ناخواسته (مخاطرات ایمنی) و همچنین آسیب‌پذیری در برابر حملات (مخاطرات امنیتی) از ابتدا تا پایان چرخه عمر سیستم هوش مصنوعی باید پیشگیری، رفع و حذف شوند تا ایمنی و امنیت انسانی و زیستی تضمین شود. ◇ فرایندهای الگوریتمی در سیستم‌های هوش مصنوعی باید نسبت به نقض حریم خصوصی افراد ارزیابی شوند و رویکرد «حریم خصوصی از طریق طراحی» را مورد توجه قرار دهند.
۳	خودمختاری (با ۷ رهنمود)	<ul style="list-style-type: none"> ◇ واگذاری امکان تصمیم‌گیری درباره انسان‌ها به سیستم‌های هوش مصنوعی باید با آگاهی از چگونگی و سپس پذیرش این فرایند مبتنی بر انتخاب خود این افراد باشد. ◇ سیستم‌های هوش مصنوعی باید با احترام به خودمختاری افراد توسعه و استقرار یابند که هدف از آن، افزایش کنترل افراد بر زندگی و محیط پیرامونی خود است. ◇ سیستم‌های هوش مصنوعی باید به افراد امکان دهند اهداف اخلاقی خود و درک شخصی خود از یک زندگی ارزشمند را تحقق بخشند. ◇ سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که سبک زندگی خاصی را (به‌طور مستقیم یا غیرمستقیم) به افراد تحمیل نکنند. ◇ توسعه و استفاده از سیستم‌های هوش مصنوعی نباید به هنگام ضرورت تصمیم‌گیری، به کاهش مسئولیت انسان‌ها منجر شود. ◇ در سیستم‌های هوش مصنوعی هرگز نباید از قدرت خودگردانی خود برای آسیب زدن، نابود کردن یا فریب دادن انسان‌ها تعبیه شود. ◇ افراد به دلایل تصمیم‌های سیستم‌های هوش مصنوعی که بر حقوق و آزادی‌های آنان تأثیر می‌گذارند باید دسترسی داشته باشند.
۴	عدالت (با ۱۰ رهنمود)	<ul style="list-style-type: none"> ◇ سیستم‌های هوش مصنوعی باید به افراد امکان دهد از ظرفیت‌های ذهنی و جسمی خود استفاده کنند. ◇ استفاده از سیستم‌های هوش مصنوعی نباید به افزایش استرس، اضطراب، یا احساس آزار از محیط دیجیتال برای افراد منجر شود. ◇ سیستم‌های هوش مصنوعی نباید تهدیدی برای روابط اخلاقی و احساس رضایت‌بخش میان انسان‌ها باشد. ◇ سیستم‌های هوش مصنوعی باید با هدف همکاری با انسان‌ها در وظایف پیچیده توسعه و استقرار یابند و کار گروهی میان انسان‌ها را تقویت کنند. ◇ سیستم‌های هوش مصنوعی باید به گونه‌ای طراحی و آموزش داده شوند که باعث ایجاد، تقویت یا بازتولید تبعیض، بر اساس تفاوت‌های اجتماعی (جنسی، قومی، فرهنگی، یا مذهبی) و دیگر موارد) نشوند.

ردیف اصل اخلاقی رهنمودهای اخلاقی مرتبط

- ◇ سیستم‌های هوش مصنوعی نباید به گونه‌ای توسعه و استقرار یابند که جایگزین افراد در وظایفی شوند که نیازمند روابط باکیفیت انسانی است.
 - ◇ سوگیری‌های تبعیض آمیز موجود در مجموعه داده‌هایی که برای آموزش و استقرار سیستم‌های هوش مصنوعی استفاده می‌شوند باید در سریع‌ترین زمان ممکن شناسایی، گزارش، و پالایش شوند.
 - ◇ سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که عدالت اجتماعی را ارتقا دهند تا از انصاف و عدم تبعیض در هر شکلی مطابق با قوانین بین‌المللی محافظت کنند.
 - ◇ سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که در طول چرخه عمر خود از تقویت یا استمرار کاربردها و نتایج تبعیض آمیز میان انسان‌ها اجتناب کنند.
 - ◇ در سطح بین‌المللی، کشورهای پیشرفته از نظر فناوری مسئولیت همبستگی با کشورهای کمتر پیشرفته را دارند تا اطمینان حاصل شود که منافع سیستم‌های هوش مصنوعی به گونه‌ای عادلانه به اشتراک گذاشته شود.
- ۵ توضیح‌پذیری (با ۱۰ رهنمود)
- ◇ هرگونه دخالت سیستم‌های هوش مصنوعی در تصمیم‌گیری‌های حیاتی باید شامل توضیحی رضایت‌بخش و قابل بررسی توسط یک مرجع انسانی معتبر باشد.
 - ◇ هر فردی که از خدمات مبتنی بر سیستم‌های هوش مصنوعی استفاده می‌کند باید امکان آگاهی از این موضوع را داشته باشد که آیا تصمیمی مربوط به او یا تأثیرگذار بر او توسط یک سیستم هوش مصنوعی اتخاذ شده است یا خیر.
 - ◇ هر فردی که از خدمات مبتنی بر چت‌بات‌ها استفاده می‌کند باید امکان تشخیص این موضوع را داشته باشد که آیا با یک سیستم هوش مصنوعی تعامل دارد یا خیر.
 - ◇ سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که معیارهای فهم‌پذیری، توجیه‌پذیری، و دسترسی‌پذیری نسبت به تصمیم‌گیری‌های خود را تحت نظارت، بررسی، و کنترل انسانی و دموکراتیک تضمین کنند.
 - ◇ سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که در صورت ایجاد آسیب به هر موجود صاحب شعوری، قابلیت ردیابی و کشف علت اصلی منشأ آن را داشته باشد.
 - ◇ سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که تصمیم‌گیری‌های آن‌ها از شفافیت و پیش‌بینی‌پذیری مناسبی برخوردار باشد تا کاربران در صورت لزوم و مبتنی بر دلایلی اخلاقی در عملکرد آن‌ها مداخله کرده یا آن‌ها را متوقف کنند.
 - ◇ فرایندها و مجموعه داده‌هایی که منجر به تصمیم‌گیری سیستم‌های هوش مصنوعی می‌شوند (از جمله مراحل جمع‌آوری داده‌ها، برجسب‌گذاری داده‌ها و الگوریتم‌های مورد استفاده) باید به بهترین شکل ممکن مستندسازی شود تا امکان ردیابی و ارتقای شفافیت در

ردیف	اصل اخلاقی	رهنمودهای اخلاقی مرتبط
		تصمیم‌های آن فراهم گردد.
◇		سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که اطلاعاتی درباره میزان تأثیرگذاری آن‌ها بر فرایند تصمیم‌گیری سازمانی، انتخاب‌های طراحی سیستم، و دلایل استفاده از آن در دسترس باشد.
◇		سیستم‌های هوش مصنوعی باید به گونه‌ای توسعه و استقرار یابند که قابلیت‌ها و محدودیت‌های آن‌ها به شکلی متناسب با کاربرد مورد نظر به متخصصان یا کاربران نهایی منتقل شود.
◇		آگاهی عمومی و فهم سیستم‌های هوش مصنوعی و ارزش داده‌ها باید از طریق آموزش‌های عمومی و قابل دسترس، مشارکت شهروندی، مهارت‌های دیجیتال، آموزش اخلاق هوش مصنوعی، سواد رسانه‌ای و اطلاعاتی و دیگر آموزش‌های مقتضی تقویت شود.

۷. نتیجه‌گیری

همان‌گونه که اشاره شد، هدف این مقاله معرفی اصول و تدوین رهنمودهای اخلاقی با مطالعه هفت سند ملی و بین‌المللی در حوزه اخلاق هوش مصنوعی بود. از آنجا که اصول و رهنمودهای اخلاقی در حوزه‌های متعدد اخلاق کاربردی (اخلاق پزشکی، اخلاق زیستی، اخلاق پژوهش، و مواردی دیگر) برآمده از تجربه‌ها و معرفت بشری است، به نظر می‌رسد که همراه با توسعه و استقرار فراگیر و جهانی سیستم‌های هوش مصنوعی، اصول و رهنمودهای اخلاق هوش مصنوعی نیز فارغ از مرزها و گستره‌های جغرافیایی برای تمام ساکنان زمین قابل استفاده و بهره‌برداری است.

از این‌رو، برای دستیابی به این هدف، پنج اصل اخلاقی یکپارچه «فلوریدی و کولنز» مبنای گردآوری و بررسی رهنمودهای اخلاق هوش مصنوعی قرار گرفت. این پنج اصل اخلاقی عبارت‌اند از: خیرخواهی، عدم آسیب‌رسانی، خودمختاری، عدالت، و توضیح‌پذیری. در ادامه، هر یک از این اصول به‌طور دقیق تعریف و قلمرو مفهومی آن‌ها برای پرهیز از فروکاست به اصل اخلاقی دیگری مشخص گردید. پس از اصول اخلاقی پنج‌گانه، رهنمودهای هفت سند اخلاق هوش مصنوعی را ذیل این اصول، بررسی و به‌ترتیب در سه مرحله به گردآوری، پالایش، و اعتباربخشی آن‌ها پرداخته شد. در مرحله اول، رهنمودهای اخلاقی اسناد مذکور استخراج و فهرست شدند. در مرحله دوم، رهنمودهای گردآوری‌شده یکدست‌سازی، موارد تکراری حذف، و رهنمودهای جزئی‌تر به نفع رهنمودهای کلی‌تر کنار گذاشته شد. در مرحله سوم، برای ارزیابی روایی یافته‌ها

و دستاورد پژوهش، رهنمودهای اخلاقی پالایش‌شده ذیل اصول اخلاقی پنج‌گانه توسط گروهی کانونی از متخصصان هوش مصنوعی، فلسفه و اخلاق فناوری و علم اطلاعات نقد و بررسی گردید. این بررسی‌ها تا زمان دستیابی به اجماعی نسبی میان متخصصان گروه کانونی به‌منظور جمع‌بندی و فهرست رهنمودهای پیشنهادی ادامه یافت. در نهایت، اصول اخلاقی خیرخواهی با ۷ رهنمود، عدم آسیب‌رسانی با ۱۲ رهنمود، خودمختاری با ۷ رهنمود، عدالت با ۱۰ رهنمود، و توضیح‌پذیری با ۱۰ رهنمود، دستاورد این پژوهش بود. در پایان لازم است به سه نکته مهم اشاره شود: اول اینکه تهیه و تدوین رهنمودهای اخلاق هوش مصنوعی، به‌دنبال ارائه راه‌حل نهایی و قطعی برای مسائل و دوراهی‌های اخلاقی بالقوه و پیچیده‌ای نیست که توسعه و استقرار سیستم‌های هوش مصنوعی بر سر راه بازیگران درگیر در چرخه حیات چنین سیستم‌هایی قرار می‌دهد. این اصول و رهنمودهای اخلاقی، در واقع، راهنمایی کلی و فراگیر است که به بازیگران درگیر در این چرخه حیات، چارچوبی برای ملاحظات اخلاقی موجه و تأملات استدلالی منسجم برای هر یک از چالش‌های متعدد اخلاقی فراهم می‌آورد. به بیان دیگر، اصول و رهنمودهای اخلاقی در این مقاله، نقش یک قطب‌نما را در مواجهه با مسائل و دوراهی‌های اخلاقی ناشی از توسعه و استقرار سیستم‌های هوش مصنوعی بازی می‌کند، نه اینکه نقشه کامل راه باشد.

دوم اینکه گروه کانونی این پژوهش، رهنمودهای مورد بحث را با رویکرد اخلاق حرفه‌ای و تأکید بر قلمرو مخاطبان اصلی آن‌ها که شامل طراحان، متخصصان و توسعه‌دهندگان سیستم‌های هوش مصنوعی است، مبنای مباحث نقادانه خود قرار دادند. اما در صورت‌بندی نهایی تلاش شده است، این رهنمودهای اخلاقی، تمامی کنشگران و عامل‌های انسانی را که در چرخه حیات این سیستم‌ها و توسعه و استقرار آن‌ها نقش‌آفرینی دارند، مخاطب قرار دهد. بی‌تردید، نقش طراحان، متخصصان و توسعه‌دهندگان سیستم‌های هوش مصنوعی برای توجه به رهنمودهای اخلاقی پیشنهادی به‌مراتب گسترده‌تر و مسئولانه‌تر است.

سوم اینکه در بررسی‌های گروه کانونی اخلاق در هوش مصنوعی، اصل عدم آسیب‌رسانی همچون بررسی اولیه هفت سند اخلاق هوش مصنوعی، از بیشترین رهنمودهای اخلاقی برخوردار شد. این موضوع، اهمیت عدم آسیب‌رسانی سیستم‌های هوش مصنوعی به جامعه انسانی و محیط زیست را نشان می‌دهد. از این رو، مطابق با رهنمودهای اخلاق هوش مصنوعی به نظر می‌رسد سیستم‌های هوش مصنوعی در طول چرخه عمر خود و در

فرایند توسعه و استقرارشان در سرتاسر کره زمین، پیش و بیش از هر دستاوردی، باید عدم آسیب رساندن به محیط زیست و موجودات صاحب شعور سکونت یافته در کره زمین را هدف نهایی خود قرار دهند.

فهرست منابع

بهاری، محمدرضا. ۱۳۹۸. *جادوی هوش مصنوعی (مجموعه‌ای درباره هوش مصنوعی)*. تهران: مؤسسه پژوهشی مطالعات فرهنگی و اجتماعی.

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

References

- Abdakhodaei, Z., N. Touhidi, N. Karimi, & M. Barati. 2021. *White Intelligence: From Machine Ethics to Ethical Machines*. Tehran: Institute for Cultural and Social Studies.
- Bahari, Mohammadreza. 2019. *The Magic of AI (A collection on artificial intelligence)*. Tehran: Institute for Cultural and Social Studies.
- Bartneck, C., C. Lütge., A. Wagner., and S. Welsh. 2021. An Introduction to Ethics in Robotics and AI. In *SpringerBriefs in Ethics*. Cham: Springer Nature.
- Boddington, P. 2017. *Towards a Code of Ethics for Artificial Intelligence*. 1st ed. Cham: Springer Publishing Company.
- Boddington, P. 2022. *AI Ethics: Textbook (Artificial Intelligence: Foundations, Theory, and Algorithms)*. Cham: Springer Publishing Company.
- Coeckelbergh, M. 2020. *AI Ethics*. Cambridge, Massachusetts. USA: The MIT Press.
- Ethically aligned design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. 2016. Retrieved from https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf. (accessed May 5, 2024)
- European Group on Ethics in Science and New Technologies. 2018. *Statement on artificial intelligence, robotics and autonomous systems*. Brussels: European Commission.
- Floridi, L. 2023. *The Ethics of Artificial Intelligence- Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.
- _____ & J. Cows. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1 (1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Future of Life Institute. 2017. *Asilomar AI Principles*. <https://futureoflife.org/ai-principles/> (accessed May 5, 2024)
- Gordon, J. & C. B. Wrenn. 2020. Ethics of artificial intelligence. In *The Internet Encyclopedia of Philosophy*. <https://iep.utm.edu/ethics-of-artificial-intelligence/>. (accessed May 5, 2022).
- High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (accessed May 5, 2024)

- House of Lords Select Committee on Artificial Intelligence. 2018. AI in the UK: Ready, willing and able? UK Parliament. <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf> (accessed May 5, 2024)
- IEEE. 2017. Ethically aligned design: A vision for prioritizing human well-being with artificial intelligence and autonomous systems (1st ed.). <https://ethicsinaction.ieee.org/> (accessed May 5, 2024)
- Mazzi, F. & L. Floridi (eds.). 2023. The Ethics of Artificial Intelligence for the Sustainable Development Goals. Cham: Springer Verlag.
- Müller, V. C. 2012. Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction. *Cognitive Computation* 4 (3): 212–215.
- Müller, V. C. 2020. Ethics of Artificial Intelligence and Robotics. In *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/>. (accessed May 5, 2022).
- Partnership on AI. 2016. Tenets. <https://partnershiponai.org/tenets/> (accessed May 5, 2022)
- Stahl, B., D. Schroeder, & R. Rodrigues. 2022. Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges. Cham: Springer Verlag.
- Technology Assessment of Digitisation Research Team. 2018. *Policy paper on the Asilomar principles on artificial intelligence*. Berlin: Office of Technology Assessment at the German Bundestag (TAB).
- UNESCO. 2021. Recommendation on the ethics of artificial intelligence. <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence> (accessed May 5, 2024)
- Université de Montréal. 2018. Montreal Declaration for Responsible AI. <https://www.montrealdeclaration-responsibleai.com/> (accessed May 5, 2024)

علیرضا نقه‌الاسلامی

در سال‌های ۱۳۸۶ و ۱۳۹۲، کارشناسی ارشد و دکتری تخصصی خود را در رشته فلسفه علم از واحد علوم و تحقیقات دانشگاه آزاد اسلامی تهران دریافت کرد. وی همکاری علمی و پژوهشی خود را به‌عنوان عضو هیئت علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) از سال ۱۳۹۴ آغاز کرد و هم‌اکنون استادیار گروه اخلاق و حقوق اطلاعات در پژوهشکده جامعه و اطلاعات این پژوهشگاه است. فلسفه و اخلاق فناوری اطلاعات، حوزه مطالعات علم و فناوری، اخلاق پژوهش در فضای مجازی و فلسفه فناوری از جمله علایق پژوهشی وی است.

