مطالعات اخلاقی
سال اول • شماره سوم • پاییز ۱۴۰۳
Journal of Ethical Studies
Vol. 1, No. 3, Autumn 2024, Issue 3

# Artificial Intelligence Systems and Ethical Challenges

**Sayyed Nasir Ahmad Hossaini** *

**Received: 2025/09/05** | **Accepted: 2025/11/04**

## Abstract

Original Research

Given the increasing penetration of artificial intelligence (AI) into various aspects of life, this technology has provided unprecedented opportunities for improving the quality of life and solving complex problems. However, it has also introduced new ethical challenges that require careful examination. The question is: what ethical challenges does AI face? This paper examines the ethical challenges associated with the development and deployment of artificial intelligence systems. Relying on the concept of "trust" in the acceptance of AI systems, the author explores its relationship with ethical principles. Trust in AI is challenging due to its inherent complexity. Users only trust systems when they are assured that they are beneficial, safe, and ethically aligned. Furthermore, the principles of autonomy and justice are crucial in AI decision-making. This technology must make decisions in critical situations that are consistent with ethical values and human rights. Additionally, explainability in the performance of AI systems is essential for building trust and ensuring accountability. This paper, which examines data using a descriptive-analytical method, seeks to analyze significant ethical challenges in AI and provide suggestions for the responsible development and deployment of this technology. The goal is to foster constructive discourse on AI ethics and develop appropriate ethical frameworks for it.

## Keywords

AI Ethics, Ethical Challenges, Trust, Non-maleficence, Beneficence, Autonomy, Explainability.

* Assistant Professor, Al-Mustafa International University, Qom, Iran. | sna.hossaini@gmail.com

Hossaini, S. N. (2024). Artificial Intelligence Systems and Ethical Challenges. *Journal of Ethical Studies, 1*(3), 67-94.
**doi: 10.22091/jes.2025.14121.1038**

مطالعات اخلاقی
سال اول • شماره سوم • پاییز ۱۴۰۳
Journal of Ethical Studies
Vol. 1, No. 3, Autumn 2024, Issue 3

# سامانه‌های هوش مصنوعی و چالش‌های اخلاقی

سید نصیر احمد حسینی *

تاریخ دریافت: ۱۴۰۴/۰۶/۱۴ | تاریخ پذیرش: ۱۴۰۴/۰۸/۱۳

علمی - پژوهشی

## چکیده

نظر به نفوذ فزایندهٔ هوش مصنوعی در جنبه‌های مختلف زندگی، این فناوری فرصت‌های بی‌نظیری برای بهبود کیفیت زندگی و حلِّ مسائل پیچیده فراهم کرده است. با این حال، چالش‌های اخلاقی جدیدی نیز به همراه داشته که نیازمند بررسی است. پرسش این است که هوش مصنوعی با چه چالش‌های اخلاقی مواجه است؟ جستار حاضر به بررسی چالش‌های اخلاقی مرتبط با توسعه و به‌کارگیری سامانه‌های هوش مصنوعی می‌پردازد. نگارنده با تکیه بر مفهوم «اعتماد» در پذیرش سامانه‌های هوش مصنوعی، ارتباط آن را با اصول اخلاقی بررسی می‌کند. اعتماد به هوش مصنوعی به دلیل پیچیدگی ذاتی آن، چالش‌برانگیز است. کاربران تنها زمانی به سامانه‌ها اعتماد می‌کنند که مطمئن شوند آنها سودمند، ایمن و اخلاق‌مدار اند. افزون بر این، اصول خودمختاری و عدالت در تصمیم‌گیری‌های هوش مصنوعی حیاتی هستند. این فناوری باید در شرایط بحرانی، تصمیم‌هایی همسو با ارزش‌های اخلاقی و حقوق انسانی بگیرد. همچنین، توضیح‌پذیری در عملکرد سامانه‌های هوش مصنوعی برای ایجاد اعتماد و پاسخ‌گویی ضروری است. این جستار ـ که داده‌ها را به روش توصیفی-تحلیلی بررسی می‌کند ـ می‌کوشد با تحلیل چالش‌های اخلاقی مهم در هوش مصنوعی، پیشنهادهایی را برای توسعه و به‌کارگیری مسئولانه این فناوری ارائه کند. هدف شکل‌گیری گفتمانی سازنده دربارهٔ اخلاق هوش مصنوعی و تدوین چارچوب‌های اخلاقی مناسب برای آن است.

## کلیدواژه‌ها

اخلاق هوش مصنوعی، چالش‌های اخلاقی، اعتماد، عدم ضرر، سودرسانی، خودمختاری، توضیح‌پذیری.

* استادیار جامعة المصطفیٰ العالمیه، قم، ایران. | sna.hossaini@gmail.com

حسینی، سید نصیراحمد. (۱۴۰۳). سامانه‌های هوش مصنوعی و چالش‌های اخلاقی. فصلنامه مطالعات اخلاقی. ۱ (۳)، ۶۷-۹۴.
doi: 10.22091/jes.2025.14121.1038

## طرح مسئله

در عصر حاضر، شاهد پیشرفت‌های چشمگیری در حوزهٔ هوش مصنوعی هستیم. سامانه‌های هوشمند با قابلیت‌های یادگیری، تصمیم‌گیری و حل مسئله، به طور فزاینده‌ای در جنبه‌های مختلف زندگی ما نفوذ کرده‌اند؛ از خودروهای خودران و دستیاران مجازی گرفته تا سامانه‌های تشخیص پزشکی و الگوریتم‌های معاملاتی بورس. این پیشرفت‌ها، فرصت‌های بی‌نظیری را برای بهبود کیفیت زندگی، افزایش بهره‌وری و حلِّ مسائل پیچیده فراهم می‌کنند. با این حال، توسعه و به‌کارگیری این سامانه‌ها، چالش‌های اخلاقی جدید و مهمی را نیز به همراه داشته است که نیازمند توجه جدّی و بررسی دقیق هستند. یکی از مهم‌ترین این چالش‌ها، چالش اخلاقی است. بنابراین، پرسش اصلی این جستار این است که هوش مصنوعی با چالش‌های اخلاقی مواجه است؟ در پی آن، پرسش‌های دیگر قابل طرح است: «اعتماد»، چه نقشی در پذیرش سامانه‌های هوش مصنوعی دارد؟ چگونه می‌توان اطمینان حاصل کرد که سامانه‌های هوش مصنوعی، تصمیم‌های مطابق با اصول و ارزش‌های اخلاقی می‌گیرد؟ آیا الگوریتم‌های هوش مصنوعی می‌توانند بی‌طرف و عادلانه عمل کنند یا ممکن است تحت تأثیر داده‌های آموزشی مغرضانه، رفتارهای تبعیض‌آمیز از خود نشان دهند؟

نگارنده در این جستار می‌کوشد به پرسش‌های بالا پاسخ دهد. نخست، دربارهٔ نقش مفاهیم اعتماد در پذیرش سامانه‌های هوش مصنوعی و سپس دربارهٔ پیوند این مفاهیم با برخی اصول اخلاقی مانند عدم‌ضرر، سودرسانی، خودمختاری، عدالت و توضیح‌پذیری بحث می‌شود. تلاش می‌شود که ابعاد مختلف این مسائل روشن شود. هدف این است که ضمن تحلیلی از چالش‌های اخلاقی مهم در حوزهٔ هوش مصنوعی، پیشنهادهایی را برای توسعه و به‌کارگیری مسئولانه این فناوری ارائه کند. نگارنده امیدوار است که این جستار به شکل‌گیری گفتمانی سازنده دربارهٔ اخلاق هوش مصنوعی و تدوین چارچوب‌های اخلاقی مناسب برای این حوزه مدد رساند.

## ۱. اعتماد کاربران به هوش مصنوعی

«اعتماد» چیست و چه نقشی در روابط میان انسان‌ها و هوش مصنوعی دارد؟ برخی پژوهشگران اعتماد را این گونه تعریف می‌کنند: اعتماد «نگرشی است که به عامل کمک می‌کند تا در شرایط نامطمئن و آسیب‌پذیر به اهداف خود دست یابد» (Lee & See, 2004, pp. 51). واژهٔ «اعتماد»[1] در زمینه‌های مختلفی به کار می‌رود و به کیفیت روابط بین افراد یا گروه‌ها اشاره دارد. اعتماد شخصی به

1. trust

خدمتکارش و به هوش مصنوعی، نمونه‌هایی از اعتماد هستند. همان‌گونه که در تعریف بالا آمده، در روابط انسانی، اعتماد همواره با درجاتی از آسیب‌پذیری همراه است. برای مثال، فروش اطلاعات کاربران توسط شرکت گوگل، نمونه‌ای از این آسیب‌پذیری هستند. بی‌اعتمادی به دیگران، سبب می‌شود انسان از سپردن امور مهم به آنها خودداری کند.

کاربران باید به سامانه‌های هوش مصنوعی اطمینان داشته باشند، اما پیچیدگی این فناوری و وابستگی آن به سامانه‌های گسترده‌تر، اعتماد کاربران به عملکرد و امنیت آنها را دشوار می‌سازد (Boddington, 2023, pp. 55)؛ زیرا مشخص نیست رُبات‌ها و سامانه‌های هوش مصنوعی، در تصمیم‌گیری خود استقلال دارند یا نه. همچنین انتظار رفتار ارزشی از رُبات‌ها و هوش مصنوعی چه‌بسا بی‌فایده یا حتی گمراه کننده باشد. باری، اعتماد به هوش مصنوعی پیچیده‌تر از اعتماد به انسانهاست.

«اعتماد» نقش کلیدی در ایجاد و حفظ پیوندهای اجتماعی حرفه‌ای دارد. به دلیل گسترش فزایندهٔ شبکه‌های اجتماعی و کسب وکارهای دیجیتال، اعتماد به فضای مجازی و سامانه‌های خودکار نیز برای پذیرش کاربر بسیار مهم است. مردم از سامانه‌های اعتمادناپذیر استفاده نمی‌کنند و محصولات شرکت‌های نامعتبر را نمی‌خرند. برای مثال، شرکت‌ها در ایالات متحده موظفند که مطابق با قانون «پاتریوت»[1] عمل کنند که به دولت اجازه دسترسی به داده‌های ذخیره شده در رایانه‌های ابری[2] را می‌دهد. مشتریان اروپایی ممکن است از اعطای چنین امتیازی به دولت ایالات متحده احساس ناخرسندی کنند. پس از اینکه دادگستری اروپا[3] توافق‌های طولانی‌مدت میان ایالات متحده و اتحادیه اروپا را با عنوان «بندر امن»[4] در سال ۲۰۱۵ لغو کرد[5]، چندین شرکت ارائه دهنده انبارش ابری[6] برای بازگشت

---

۱. قانون پاتریوت ایالات متحده (USA PATRIOT Act) قانون فدرال بسیار مهم و بحث‌برانگیز است که در واکنش به حادثه تروریستی ۱۱ سپتامبر ۲۰۰۱ تصویب شد. عنوان «USA PATRIOT» مخفف عبارت Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism به معنای «قانون اتحاد و تقویت آمریکا با ارائه ابزارهای مناسب مورد نیاز برای رهگیری و جلوگیری از دهشت‌افکنی (تروریسم)» است. این قانون – پس از حملات ۱۱ سپتامبر و برای تقویت امنیت ملی آمریکا تصویب شد.

2. cloud computers

3. European Court of Justice

4. US-EU Safe Harbor

۵. توافق‌نامه «بندر امن» بین ایالات متحده و اتحادیه اروپا (US-EU Safe Harbor) با هدف تسهیل انتقال داده‌های شخصی و حفاظت از حریم خصوصی ایجاد شد. شرکت‌های آمریکایی ملزم به رعایت هفت اصل حریم خصوصی بودند. با این حال، دادگاه دادگستری اروپا این توافق‌نامه را به دلیل عدم حفاظت کافی در برابر نظارت دولتی آمریکا لغو کرد. لغو «بندر امن» منجر به ایجاد «سپر حریم خصوصی» (Privacy Shield) شد که آن هم به دلیل چالش‌های قانونی لغو شد. در حال حاضر، شرکت‌ها با قوانین سخت‌گیرانه‌تری مانند مقررات عمومی حفاظت از داده‌ها (GDPR) مواجه هستند.

6. cloud storage company

اعتماد مشتریان اروپایی خود، مراکز داده‌ای در اروپا تأسیس کردند (Bartneck et al, 2021, pp. 28).

«فرهنگ» نقش بسزایی در میزان اعتماد افراد به سامانه‌های هوش مصنوعی و رُبات‌ها دارند. این تأثیر ابعاد مختلفی از تعامل انسان و ماشین را در بر می‌گیرد. تحقیقات نشان می‌دهد که تفاوت‌های فرهنگی، نگرش افراد نسبت به ربات‌ها را شکل می‌دهد (Haring et al. 2014a). به دیگر سخن، اینکه یک فرد در چه فرهنگی پرورش یافته، دیدگاه او را در مورد ربات‌ها، از جمله میزان پذیرش، ترس یا کنجکاوی نسبت به آنها، تحت تأثیر قرار می‌دهد. مطالعات میان‌فرهنگی همچنین تفاوت‌هایی را در میزان مثبت‌نگری نسبت به فناوری‌های جدید، از جمله هوش مصنوعی و رباتیک، نشان می‌دهد. این مثبت‌نگری به‌طور مستقیم بر اعتماد افراد به این سامانه‌ها اثر می‌گذارد (Haring et al. 2014b). برای مثال، در برخی فرهنگ‌ها، پذیرش فناوری‌های جدید و اعتماد به آنها بیشتر از دیگر فرهنگ‌هاست که این امر می‌تواند برخاسته از عواملی مانند سطح توسعه‌یافتگی فناوری در آن جامعه، نگرش‌های دینی یا فلسفی غالب و یا تجربه‌های تاریخی باشد.

عوامل فرهنگی فرهنگی نقش تعیین‌کننده‌ای در میزان پذیرش توصیه‌های ربات‌ها و سامانه‌های هوش مصنوعی دارند (Wang et al. 2010). در فرهنگ‌هایی که احترام به مراجع خارجی به ویژه متخصصان یا دانشوران پررنگ است، افراد تمایل بیشتری به پیروی از دستورات ماشینی نشان می‌دهند. در مقابل، در فرهنگ‌های فردگرا که بر استقلال رأی و تحلیل شخصی تأکید می‌شود، افراد معمولاً توصیه‌های ربات‌ها را به‌دقت بررسی کرده و کمتر بدون ارزیابی، آنها را می‌پذیرند. باری، تأثیر فرهنگ بر رفتار افراد و چگونگی تعامل آنها با فناوری‌های جدید مانند هوش مصنوعی و رباتیک و میزان اعتماد به آنها انکارناپذیر است. درک این تأثیرات فرهنگی برای طراحی و توسعهٔ سامانه‌های هوش مصنوعی و رباتیک به منظور پذیرش و بهره‌برداری مؤثر از آنها بسیار حائز اهمیت است.

اعتماد به ماشین و سامانه‌های هوش مصنوعی، پدیده‌ای چند بُعدی است که می‌توان عوامل و ابعاد آن را به «کارکردی» و «اخلاقی» تقسیم کرد.

## ۲. جنبه‌های کارکردی اعتماد به هوش مصنوعی

اعتماد به سامانه‌های خودکار، فرایندی پیچیده است که تحت تأثیر عوامل گوناگونی قرار دارد. عناصر کارکردی اعتماد به سامانه‌های خودکار و ماشین‌ها، عمدتاً بر قابلیت‌ها و عملکرد فنی آنها مانند دقت، سرعت و کارایی متمرکز است. کاربران سطح اعتماد خود به ماشین‌ها را بر اساس معیارهای مختلفی تنظیم می‌کنند که یکی از مهم‌ترین آنها قابلیت اطمینان سامانه است (Lee & See, 2004). پژوهش‌های متعددی به بررسی عوامل مختلفی پرداخته‌اند که بر اعتماد به دستگاه‌های

خودکارسازی (اتوماسیون)[1] تأثیر می‌گذارند (Hancock et al., 2011). عوامل مرتبط با عملکرد سامانه عبارتند از قابلیت اتکا به سامانه، نرخ هشدارهای نادرست، شفافیت عملکرد و پیچیدگی وظایف محوله به سامانه. نتایج این مطالعات نشان می‌دهد که کارایی سامانه و قابلیت اطمینان آن، از جمله عوامل کلیدی و غالب در پذیرش فناوری‌های خودکار شناخته می‌شوند. مردم در مواجهه با ماشین‌های بالقوه خطرناک با احتیاط رفتار می‌کنند، بویژه زمانی که آسیب‌های جانی یا صدمات جدّی و دائمی وجود دارد. این نگرانی‌ها سبب کاهش اعتماد به چنین فناوری‌هایی می‌شود.

## ۳. جنبه‌های اخلاقی اعتماد به هوش مصنوعی

اعتماد به ماشین صرفاً به توانایی انجام وظایف و عملکرد فنی آن محدود نمی‌شود، بلکه دارای ابعاد اخلاقی مهمی است. کاربران تنها به ماشینی که کارش را به درستی انجام می‌دهد اعتماد نمی‌کنند، بلکه به ماشینی اعتماد می‌کنند که مطابق با اصول اخلاقی عمل می‌کند و رفتار آن پیش‌بینی‌پذیر و توجیه‌پذیر است. پژوهشگران شمار اصول اخلاقی‌ای را که در حوزه هوش مصنوعی و رباتیک به کاربسته می‌شود در آثار خود متفاوت ذکر کرده‌اند (Boddington, 2023, pp. 45-60; Floridi, 2023, pp. 57-60)، اما اصولی که روی آن اتفاق نظر وجود دارد و در عین حال، اهمیت بیشتری دارد، پنج اصل است که در اسناد کشورهای مختلف روی آنها تأکید شده است. سند یا رهنمود «اصول اخلاقی اروپا برای هوش مصنوعی»، که در سال ۲۰۱۸ منتشر شد، پنج اصل مهم را برای اخلاق هوش مصنوعی پیشنهاد می‌کند (Floridi et al. 2018, pp. 695-700; Harasimiuk & Braun, 2021, pp. 66-69). این اصول عبارتند از سودرسانی، زیان‌نرساندن، خودمختاری، عدالت و توضیح‌پذیری که با مفاهیم اعتماد و انصاف مرتبط هستند. این اصول اخلاقی ـ که ریشه در حقوق اساسی دارند ـ باید رعایت شوند تا اطمینان حاصل شود که سامانه‌های هوش مصنوعی قابل اعتماد هستند. در ادامه، دربارهٔ این اصول و ارتباط آنها با اعتماد در ادامه، بررسی می‌شود.

### ۳. ۱. بهره‌گیری از اصل سودرسانی در هوش مصنوعی

اصل سودمندی[2] یا سودرسانی در هوش مصنوعی بیان می‌کند که این فناوری باید برای مردم و جامعه مفید باشد. این یکی از اصول بنیادین اخلاق زیستی و اخلاق پزشکی است که بر اساس آن، منافع

1. automation
2. Beneficence

درمان باید بیشتر از زیان‌های احتمالی آن باشد (Beauchamp & Childress, 2012, pp. 202). بر مبنای این اصل، فواید فناوری‌های دیجیتال باید بیشتر از زیان‌ها و خطرات احتمالی آنها باشد. سامانه‌های هوش مصنوعی برای اعتمادپذیری و بهبود کیفیت زندگی، باید اصول اخلاقی را رعایت کنند.

هوش مصنوعی می‌تواند در زمینه‌های مختلفی سودمند باشد که به برخی موارد اشاره می‌کنیم. هوش مصنوعی می‌تواند با تحلیل داده‌های ترافیکی به بهبود جریان ترافیک و کاهش زمان سفر کمک کند. همچنین، می‌تواند در تشخیص بیماری‌ها، پیشگیری و درمان زوال عقل، و ارائهٔ خدمات پزشکی از راه دور مؤثر باشد. در زمینهٔ انرژی، می‌تواند به بهبود پایداری و بهره‌وری مصرف آب کمک کند. در حفظ محیط زیست، می‌تواند در حفاظت از گونه‌های در معرض خطر و مقابله با تغییرات آب و هوایی نقش داشته باشد. در آموزش، می‌تواند به عنوان دستیار معلمان و دانش‌آموزان عمل کند و یادگیری شخصی را فراهم کند. در نهایت، در امنیت و عدالت اجتماعی، می‌تواند به پیشگیری از جُرم و تشخیص اخبار جعلی کمک کند.

### ۳. ۲. کاربرد اصل «زیان‌نرساندن» در هوش مصنوعی

اصل «زیان‌نرساندن»[1] یا «عدم ضرر» اصل کلی دیگر از اصول بنیادین اخلاق زیستی و پزشکی است (Beauchamp & Childress, 2012, pp. 150-151) اصل عدم ضرر یا پیشگیری از آسیب جزو اصول اخلاقی اتحادیه اروپا در توسعه و استفاده از هوش مصنوعی قابل اعتماد نیز به شمار می‌رود. این اصل نه‌تنها از لحاظ اخلاقی حیاتی است که از نظر عملی نیز بسیار مهم است؛ زیرا سامانه‌های هوش مصنوعی آسیب‌زا هم خطری برای افراداند و هم تهدیدی برای اعتبار شرکت‌های توسعه‌دهنده. بنا به این اصل، الگوریتم‌های هوش مصنوعی باید از تبعیض، دستکاری و ایجاد پروفایل‌های منفی اجتناب کنند و باید از گروه‌های آسیب‌پذیر مانند کودکان و مهاجران محافظت نمایند (Coeckelbergh, 2020, pp. 152-153). در بحث اعتماد و عدالت، این اصل شامل مواردی مانند زورگویی و سخنان نفرت‌آمیز نیز می‌شود که مصادیق بارز نقض اصل «زیان نرساندن» در فضای مجازی و برخط هستند.

#### ۳. ۱. ۱. زورگویی و آزار در رسانه‌های اجتماعی

زورگویی در محیط‌های آموزشی و کاری پدیده‌ای رایج و نگران‌کننده است که می‌توانند تأثیرات عمیق

1. Non-maleficence

و جبران‌ناپذیری بر سلامت روانی قُربانیان بگذارند. در برخی موارد، این آزارها به پیامدهای غم‌انگیزی مانند افسردگی شدید، اضطراب و حتی خودکشی انجامیده است. هرچند چنین رفتارهایی پیش از ظهور فناوری‌های دیجیتال نیز وجود داشته، نظرسنجی‌ها نشان می‌دهد که گسترش شبکه‌های اجتماعی و فضای مجازی ابعاد جدید و نگران‌کننده‌ای به این پدیده بخشیده است (Marshal, 2017). زورگویی در فضای برخط به شکل‌های مختلفی مانند ارسال پیام‌های آزاردهنده، تهدیدآمیز یا اشتراک‌گذاری محتوای تحقیرآمیز ظاهر می‌شود. این نوع زورگویی، به دلیل گستردگی و سرعت انتشار در فضای مجازی، می‌تواند تأثیرات ویرانگرتری داشته باشد؛ زیرا قُربانیان ممکن است در معرض دید هزاران یا حتی میلیون‌ها نفر قرار بگیرند و احساس درماندگی و انزوا کنند.

در واکنش به این چالش فزاینده، بسیاری از کشورها اقدامات قانونی را برای مقابله با زورگویی و ارعاب در رسانه‌های اجتماعی تصویب کرده‌اند. برای مثال، سوئد کشوری پیشگام در این زمینه بود که در سال ۱۹۹۳ قانونی را برای مبارزه با این پدیده به تصویب رساند. هدف از این قانون، حمایت از قُربانیان و مجازات افراد زورگو است. دیگر کشورها نیز با الهام از این اقدامات، قوانین مشابهی را وضع کرده‌اند تا از حقوق افراد در برابر آزار و اذیت محافظت کنند.

با این حال، افزون بر تصویب قوانین، آموزش و آگاهی‌بخشی به افراد دربارهٔ پیامدهای زورگویی و اهمیت احترام به دیگران، نقش کلیدی در پیشگیری از این پدیده دارد. همچنین، ایجاد فضایی امن و حمایت‌کننده در محیط‌های مختلف می‌تواند به کاهش موارد زورگویی و بهبود سلامت روانی جامعه کمک کند. در نهایت، مقابله با زورگویی نیازمند همکاری همه‌جانبهٔ دولت‌ها، نهادهای آموزشی، خانواده‌ها و خود افراد جامعه است تا بتوان به محیطی امن و عاری از آزار برای همه دست یافت.



### ۳. ۱. ۲. نفرت‌پراکنی در فضای مجازی

نفرت‌پراکنی و سخنان نفرت‌انگیز به مثابه معضلی جدی در سال‌های اخیر بازتاب گسترده‌ای در رسانه‌های مختلف داشته است. نفرت‌پراکنی سخنی است که تنها بر ضد افرادی خاص انجام نمی‌گیرد، بلکه می‌تواند بر ضد گروه‌ها یا بخش‌های بزرگی از جامعه ابراز شود. برای مثال، سخنان نفرت‌انگیز می‌تواند گروه‌ها را بر اساس منشأ قومی، گرایش جنسی، دین، جنسیت، هویت، ناتوانی و موارد دیگر مورد حمله قرار دهد. این حملات می‌تواند مستقیم و آشکار باشند، مانند استفاده از الفاظ توهین‌آمیز، یا غیرمستقیم و پنهان، مانند انتشار کلیشه‌های منفی و تبعیض‌آمیز. نفرت‌پراکنی، نه‌تنها به قُربانیان آن آسیب روانی می‌رساند، بلکه می‌تواند به خشونت و تبعیض در جامعه نیز بینجامد. گسترش

این نوع گفتار، فضایی از ترس و ناامنی را در جامعه ایجاد می‌کند و مانع همبستگی و تعامل سازنده بین گروه‌های مختلف می‌شود. «میثاق بین المللی حقوق سیاسی و مدنی»[1] که از سال ۱۹۷۶ لازم الاجرا شده، شامل بیانیه‌ای است که بر اساس آن «هر گونه حمایت از نفرت ملی، نژادی یا دینی که تحریک به تبعیض، دشمنی یا خشونت را در پی داشته باشد، طبق قانون ممنوع است» (ICCPR, Art, pp. 20).

از آن زمان تا کنون، قوانین متعددی بر ضد نفرت‌پراکنی در شماری از کشورها تصویب شده است. برای نمونه، بلژیک در سال ۱۹۸۱ قانونی را وضع کرد که بر اساس آن تحریک به تبعیض، نفرت یا خشونت بر ضد افراد یا گروه‌ها بر بنیاد نژاد، رنگ پوست، گروه‌هایی با منشأ خاص یا مخالفت‌های ملی یا نژادی، همه غیرقانونی بوده، طبق قانون جزای بلژیک مشمول مجازات است. کشورهای دیگری مانند کانادا، نیوزیلند و بریتانیا نیز قوانینی مشابه دارند که ممکن است در جزئیات و میزان مجازات با یکدیگر تفاوت داشته باشند (Bartneck et al, 2021, pp. 29; Gorenc, 2022, pp. 414-416).

با این حال، سطح جدیدی از قوانین مقابله با سخنان نفرت‌انگیز در سال ۲۰۱۷ در آلمان وضع شد. در این سال لایحه‌ای توسط مجلس ملی آلمان به تصویب رسید که به‌طور خاص نفرت‌پراکنی در رسانه‌های اجتماعی را جرم‌انگاری می‌کند. این قانون نیز تأکید می‌کند که شبکه‌های اجتماعی در صورتی که ظرف یک هفته به طور فعّال موفق به شناسایی و حذف محتوای نفرت‌انگیز نشوند، با پرداخت جریمه‌های بسیار سنگینی تا سقف ۵۰ میلیون یورو روبه رو خواهند شد. این قانون، با هدف مقابله با گسترش سریع و آسان سخنان نفرت‌آمیز در فضای مجازی وضع شده است. دولت آلمان امیدوار است که با اعمال این قانون سختگیرانه، شرکت‌های رسانه‌های اجتماعی را مجبور به نظارت دقیق‌تر بر محتوای منتشر شده در سکوهای خود و حذف سریع‌تر محتوای غیرقانونی کند.

این قانون که در آلمان تصویب شد، با واکنش‌های زیادی روبه رو شد و بسیاری از تحلیل‌گران آن را قانونی سختگیرانه با پیامدهای ناخواسته ارزیابی کردند. از زمان تصویب، شبکه‌های اجتماعی مانند ایکس و فیس‌بوک تلاش زیادی برای اجرای آن انجام داده‌اند و به اذعان کمیسیون اروپا در سال ۲۰۱۹، در حذف محتوای غیرقانونی موفق بوده‌اند. با این حال، اجرای این قانون مشکلاتی را به‌وجود آورده که از جمله آنها حذف ناخواسته یا تفسیر نادرست برخی مطالب است (Bartneck et al, 2021, pp. 29). به همین دلیل، این شرکت‌ها گاهی مجبور به حذف محتوایی می‌شوند که لزوماً غیرقانونی نیست، اما می‌تواند مورد انتقاد قرار گیرد. این مسئله می‌تواند به محدودیت آزادی بیان و خودسانسوری در فضای مجازی بینجامد. مشکل دیگر، عدم شفافیت در فرایند حذف محتواست.

1. International Covenant on Civil and Political Rights

کاربران اغلب نمی‌دانند که چرا محتوای آنها حذف شده و فرصت کافی برای اعتراض به این تصمیم‌ها را ندارند که این موضوع می‌تواند به نارضایتی و بی‌اعتمادی کاربران بینجامد.

با این همه، امکان اعتراض به حذف محتوا اهمیت زیادی دارد و شبکه‌های اجتماعی در حال ایجاد ساز و کارهایی برای آن هستند. این ساز و کارها به کاربران امکان می‌دهند تا در صورت حذف نابه‌جا، اعتراض کرده و موضوع دوباره بررسی شود. در نهایت، قانون مذکور نشان می‌دهد که مبارزه با نفرت‌پراکنی در فضای مجازی پیچیده و چالش‌برانگیز است و نیازمند رویکردی متوازن برای جلوگیری از محتوای غیرقانونی و محافظت از آزادی بیان است.

موضوع حد و مرز بین سخنان نفرت‌انگیز و آزادی بیان در سطوح ملّی و بین‌المللی مورد بحث است (Gorenc, 2022, pp. 418). رسانه‌های اجتماعی با افزودن جنبه‌های فرهنگی و ملی، این مباحث را پیچیده‌تر کرده‌اند. شرکت‌های بزرگ فناوری مانند فیس‌بوک و گوگل برای مقابله با نژادپرستی و تبعیض نژادی، محتوای خود را ویرایش می‌کنند. برخی این اقدام را ضروری می‌دانند و معتقدند که این شرکت‌ها باید در برابر انتشار سخنان نفرت‌انگیز مسئول باشند. در مقابل، برخی دیگر این اقدام را یورش به آزادی بیان و مردم‌سالاری تلقی می‌کنند و بر این باورند که شرکت‌های فناوری نباید محتوای منتشر شده توسط کاربران را محدود کنند. به باور این دسته، آزادی بیان حقی اساسی برای همه افراد است و هیچ نهادی نباید آن را محدود کند.

اختلاف نظرها نشان می‌دهد که تعیین حد و مرز بین سخنان نفرت‌انگیز و آزادی بیان بسیار پیچیده و حسّاس است. برای یافتن راه‌حلی مناسب، باید تعریف دقیق و مشخصی از سخنان نفرت‌انگیز ارائه شود و به زمینه و شرایط بیان آن توجه شود. یک عبارت مشابه ممکن است در شرایطی خاص، سخن نفرت‌انگیز تلقی شود، در حالی که در شرایط دیگر، نمی‌توان آن را مصداق نفرت‌پراکنی دانست. همچنین، هر گونه اقدامی در محدود کردن این سخنان باید با رعایت اصول حقوق بشر و آزادی بیان انجام شود. گفت‌وگوی آزاد و تبادل نظر در این زمینه می‌تواند به یافتن راه‌حلی جامع و مورد قبول برای همه کمک کند.

در نهایت، ایجاد تعادل بین آزادی بیان و مقابله با سخنان نفرت‌انگیز نیازمند بررسی دقیق و کارشناسانه ابعاد مختلف این موضوع و همکاری بین صاحب‌نظران، حقوق‌دانان، فعالان اجتماعی و شرکت‌های فناوری است.

اینها تنها چند نمونه از کاربردهای هوش مصنوعی در حلِّ مشکلات اجتماعی هستند. با پیشرفت روزافزون این فناوری، می‌توان انتظار داشت که در آینده شاهد کاربردهای بیشتر و مؤثرتری از آن در زمینه‌های مختلف باشیم.

### ۳. ۳. خودمختاری در هوش مصنوعی

اصل خودمختاری[1] در هوش مصنوعی به معنای احترام به اهداف و خواسته‌های افراد است. البته، خودمختاری معانی متعددی دارد (Bartneck et al, 2021, pp. 83, 94, 95). در حالی که در هوش مصنوعی و رُباتیک[2]، «خودمختاری» به توانایی سامانهٔ هوش مصنوعی یا ربات در انجام عملیاتی بدون دخالت انسان اشاره دارد، در این اینجا منظور جنبهٔ اخلاقی آن است. خودمختاری در اخلاق زیستی به حق بیماران در تصمیم‌گیری در مورد درمان خود اشاره دارد. بر اساس این اصل، آنها حق دارند خود تصمیم بگیرند که تحت درمان قرار گیرند یا نه (Beauchamp & Childress, 2012, pp. 101-105). این اصل شامل حق بیماران برای خودداری از انجام روش‌های درمانی نجات‌بخش یا اجتناب از مصرف داروهای کاهش‌دهندهٔ خطر نیز می‌شود. برای مثال، مواردی شناخته شده که در آن افراد به دلیل امتناع از دریافت انتقال خون به دلایل دینی جان خود را از دست داده‌اند. به‌طور کلی، دادگاه‌ها دریافته‌اند که والدین حق ندارند نظرات خود مانند امتناع از انتقال خون را به فرزندان خود تحمیل کنند (Woolley, 2005, pp. 715-716). با این همه، هنگامی که کودکی به سن قانونی می‌رسد، می‌تواند خود تصمیم بگیرد و از درمان خودداری کند.

به‌طور کلی، «خودمختاری» به توانایی فرد در تصمیم‌گیری اشاره دارد (Christman, 2020)، مانند انتخاب بین درآمد بیشتر یا پذیرش مخاطرات. برای مثال، شِرپاهای[3] نپالی با حمل کوله‌پشتی کوهنوردان به قلهٔ اورست، پنج برابر بیشتر از کار در مزارع درآمد دارند، اما با خطراتی مانند بهمن مواجهند. افراد باید آزاد باشند تا خطرات را بپذیرند، اما باید از عواقب آن آگاه باشند. احترام هوش مصنوعی به خودمختاری انسان نیز شامل پذیرش خطرات شخصی مانند صخره‌نوردی یا موتورسواری است (Bartneck et al, 2021, pp. 31).

خودمختاری در هوش مصنوعی و رباتیک با محدودیت‌های اخلاقی مهمی همراه است. برای مثال، اگر دانش‌آموزی از رباتی بخواهد در تقلب امتحان به او کمک کند یا فردی از ربات بخواهد که شایعات دروغینی را دربارهٔ همسایه‌اش در فضای مجازی پخش کند، رُبات نباید از این دستورات و خواسته‌ها پیروی کند. سامانه‌ها نباید در انجام اقدامات غیرقانونی یا غیر اخلاقی به افراد

---

1. Autonomy
2. robotic

۳. شرپاها (Sherpas) (به زبان رومی: shar pa) یکی از گروه‌های قومی تبتی بومی کوهستانی‌ترین مناطق نپال و منطقه خودمختار تبت هستند. اصطلاح شرپا از کلمات تبتی shar (به معنای «شرق») وpa (به معنای «مردم») گرفته شده که به منشاء جغرافیایی آنها در تبت شرقی اشاره دارد.

کمک کنند. هیچ دلیل قانع‌کننده‌ای وجود ندارد که اجازه دهد سامانه‌ها برای آسیب رساندن به دیگران مورد استفاده قرار گیرند، مگر در مواردی که دلیل موجّهی وجود داشته باشد. طراحی سامانه‌هایی که به ماشین‌ها اجازه می‌دهند به درخواست افراد آسیب برسانند، نیازمند دقت و احتیاط بسیار است. به عنوان مثال، ساخت رُبات‌های مرگبار می‌تواند عواقب جبران‌ناپذیری داشته باشد؛ زیرا ممکن است آنها قادر به تشخیص شرایط حسّاس مانند بیماری‌های روانی نباشند. بنابراین، رعایت اصول اخلاقی در طراحی و استفاده از هوش مصنوعی و رباتیک اجتناب‌ناپذیر است.

با این حال، در برخی موارد، ربات‌ها برای استفاده از زور علیه انسان‌ها طراحی می‌شوند، مانند استفاده پلیس از ربات‌ها در مواجهه با مُجرمان خطرناک. برای مثال، در سال ۲۰۱۶ در دالاس آمریکا، پلیس از ربات کنترل از راه‌دور برای انتقال مواد منفجره و کشتن مجرمی استفاده کرد که به ده افسر شلیک کرده، پنج تن را کشته و پنج تن دیگر را زخمی کرده بود (Thielmann, 2016). در این عملیات که به مرگ مُجرم انجامید، انسان‌ها کنترل کامل ربات را در دست داشتند. همچنین، در حوزه نظامی سال‌هاست از ربات‌های خشونت‌آمیز استفاده می‌شود. با این حال، به جز موارد قانونی اِعمال خشونت توسط دولت، اکثر افراد با طراحی ربات‌ها و هوش مصنوعی برای آسیب رساندن به انسان‌ها مخالفند.

مفهوم خودمختاری، که به «حکومت بر خود» تعریف می‌شود (Bunnin & Yu, 2004, pp. 63)، ابتدا به موجودیت‌های سیاسی خودگردان اشاره داشت، اما با ظهور کانت، این مفهوم با استقلال عقل عملی و آزادی فرد برای مدیریت امور خود بدون وابستگی به دیگران، از جمله دولت، پیوند خورد. این برداشت از خودمختاری فردی، به‌تدریج گسترش یافت و شامل اداره زندگی بر اساس خواسته‌ها و تمایلات شخصی شد (Formosa, 2021, pp. 597; Reath, 2005, pp. 75-76). خودمختاری با فلسفهٔ اخلاق پیوندی مستقیم دارد؛ اگر شخصی، خودمختاری نداشته باشد، مسئولیت اخلاقی نیز نخواهد داشت. پیوند میان خودمختاری و اخلاق را می‌توان به‌خوبی در نظریه وظیفه‌گرایانهٔ[1] ایمانوئل کانت[2] (۱۷۲۴-۱۸۰۴)، فیلسوف نامدار آلمانی، مشاهده کرد. این نظریه سه صورت‌بندی اصلی دارد:

صورت اول: «من هیچ گاه نباید جز این رفتار کنم تا که همچنین بتوانم اراده کنم که آیین رفتارم به قانونی عام[3] مبدّل شود» (کانت، ۱۳۹۴، ص ۴۸؛ سالیوان، ۱۳۸۰، ص ۶۲)؛ صورت دوم: «چنان رفتار کن تا بشریت را چه در شخص خود و چه در شخص دیگری همیشه به عنوان یک غایت به شمار آوری، و نه هرگز تنها همچون وسیله» (کانت، ۱۳۹۴، ص ۴۸؛ سالیوان، ۱۳۸۰، ص ۶۲)؛ صورت سوم:

1. deontological point of view
2. Immanuel Kant
3. universal law

«همهٔ آیین‌های رفتار باید به‌واسطهٔ قوانینی که خود وضع می‌کنند با مملکت ممکن غایات[1] و مملکت طبیعت[2] هماهنگ شوند» (کانت، ۱۳۹۴، ص ۱۰۶؛ سالیوان، ۱۳۸۰، ص ۶۲).

ضابطهٔ قانون همگانی می‌گوید برای توجیه اخلاقی یک عمل، باید بتوانیم قاعده اخلاقی حاکم بر آن را به عنوان قانونِ جهانی در نظر بگیریم. به دیگر سخن، اگر قاعده‌ای که بر اساس آن عمل می‌کنیم بتواند به عنوان قانون کلی برای همهٔ انسان‌ها در هر زمان قابل اجرا باشد، آنگاه آن عمل از نظر اخلاقی توجیه‌پذیر است. این اندیشه بر این فرض استوار است که اخلاق باید جهان‌شمول و بی‌طرفانه باشد. اگر رفتاری جامعه‌ای را به سوی هماهنگی، عدالت و رفاه سوق دهد، آن رفتار از نظر اخلاقی توجیه‌پذیر است. اصل انسانیت بیان می‌کند که نباید از دیگران همچون «وسیلهٔ صرف» برای اهداف خود استفاده کنیم، بلکه باید اهداف و خواسته‌های آنها را به اندازه خود ارزشمند دانسته، به آنها احترام بگذاریم. این اصل بر ارزش ذاتی انسان‌ها، احترام به خودمختاری و رفتار اخلاقی تأکید می‌کند.

کانت، مفهوم «مملکت غایات» از امر مطلق[3] را در فلسفهٔ اخلاق خود مطرح کرده است. این مفهوم بیان می‌کند: «هر ذات خردمند باید چنان عمل کند که گویی از راه آیین‌های کردارش، عضوی قانون‌گذار در مملکت همگانی غایات است» (کانت، ۱۳۹۴، ص ۱۱۰). اصول خودمختاری و مملکت غایات بر این نکته تأکید دارند که باید از همان قواعد اخلاقی پیروی کنیم که از دیگران انتظار داریم. به باور کانت، انسان‌ها به عنوان موجودات خودمختار وظیفه دارند معیارهای اخلاقی خود را عقلانی بررسی کنند و صرفاً کورکورانه از قوانین پیروی نکنند. انسان‌های خردمند دارای ارزش ذاتی هستند و باید بر اساس عقل و ارادهٔ آزاد خود، قوانین اخلاقی را تعیین و اجرا کنند.

حال، چه پیوندی میان اخلاق کانت با هوش مصنوعی است؟ با توجه به دیدگاه اخلاقی کانت، می‌توان استدلال کرد که سامانهٔ هوش مصنوعی هوشمند[4] برای اینکه بتواند به شیوه اخلاقی عمل کند باید به معنای کانتی «خودمختار» باشد. این بدان معناست که سامانه باید بتواند با استفاده از الگوریتم‌ها و معیارهای اخلاقی مناسب، تصمیم‌های اخلاقی بگیرد و دلایل تصمیم‌های خود را به شکل شفاف و قابل فهم توضیح دهد.

بسیاری صاحبنظران، مسئولیت اخلاقی[5] و تصمیم‌گیری اخلاقی[6] را در تعریف خود از چیستی

1. possible kingdom of ends
2. kingdom of nature
3. Imperative
4. artificial intelligent system
5. moral responsibility
6. moral decision making

«عامل اخلاقی»[1] ترکیب کرده‌اند (Floridi & Sanders, 2004). این دیدگاه، عامل اخلاقی را با ظرفیت پاسخ‌گویی در قبال اعمال خود برابر می‌داند. این نشان می‌دهد که اگر موجودی نتواند برای انتخاب‌های خود سرزنش یا ستایش شود، نمی‌توان او را بازیگر اخلاقی واقعی در نظر گرفت. این دیدگاه اغلب از فهم سنّتی و انسان‌محور اخلاق نشأت می‌گیرد؛ جایی که پاسخ‌گویی سنگ بنای رفتار اخلاقی است. البته، آنها به‌طور معمول مدافع تفسیر سختگیرانه از عاملیت اخلاقی هستند. از نظر آنها، عامل اخلاقی باید چندین ویژگی مؤثر، از جمله تعامل[2]، خودمختاری، انطباق‌پذیری[3] و مسئولیت اخلاقی را داشته باشد (Floridi & Sanders, 2004, pp. 352-368). آنها معتقدند که هوش مصنوعی فعلی در برآوردن معیار مسئولیت اخلاقی ناکام است؛ زیرا فاقد آگاهی، هدف، قصد و ارادهٔ آزاد است. از نظر آنها، «عامل مصنوعی می‌تواند مسئولیت اخلاقی داشته باشد اگر و تنها اگر عامل اخلاقی باشد» (I loridi & Sanders, 2004, p. 359). این بدان معناست که اگر هوش مصنوعی نتواند از نظر اخلاقی مسئول شناخته شود، نمی‌توان آن را عامل اخلاقی کامل در نظر گرفت.

برخی دیگر این دو مفهوم را از هم جدا کرده، معتقدند که هوش مصنوعی می‌تواند تصمیم‌های اخلاقی بگیرد بی آنکه مسئول تصمیم‌های خود باشد (Welsh, 2018, pp. 29-31). از نظر کانت، سامانه‌ای که برنامه‌ریزی شده تنها از قوانینی مانند سه قانون رُباتیک آسیموف[4] (۱۹۲۰-۱۹۹۲) پیروی کند[5]، عامل اخلاقی شناخته نمی‌شود.

### ۳. ۴. عدالت و انصاف در هوش مصنوعی

اصل عدالت، بیان می‌کند که هوش مصنوعی باید به‌طور عادلانه و بی‌طرفانه عمل کند. تعریف

1. ethical agent
2. interactivity
3. adaptability
4. Asimov's Three Laws of Robotics

۵. آیزاک آسیموف (Isaac Asimov) سه قانون رباتیک را پیشنهاد کرد که هدف آنها محافظت از بشریت در برابر ربات‌های بدخواه بود: الف) رُبات ممکن است به انسان آسیب نرساند یا با اقدام نکردن، باعث آسیب رساندن به انسان شود؛ ب) ربات باید از فرمان‌های انسان اطاعت کند، مگر اینکه چنین فرمان‌هایی با قانون اول مغایرت داشته باشد؛ ج) ربات باید از وجود خود محافظت کند تا زمانی که چنین حفاظتی با قانون اول یا دوم در تضاد نباشد (Turner, 2019: 2; Iphofen & Kritikos, 2019: 8)؛ هرچند آثار آسیموف در رسانه‌های عمومی بسیار شناخته شده‌اند، با انتقاد فیلسوفان روبه‌رو شده‌اند. آسیموف در پی انتقادها در نهایت، «قانون صفر» را نیز اضافه کرد: «رُبات، نباید به انسانها آسیب برساند، یا با اقدام نکردن اجازه دهد انسانها دچار آسیب شود».

«عدالت»[1] در سطح انسانی چالشی بزرگ برای هوش مصنوعی است؛ زیرا دربارهٔ نظریه‌های اخلاقی ناظر به آن، بحث‌های زیادی وجود دارد. از دوران سقراط تا امروز، مکاتب فکری مختلف تعاریف متفاوتی از عدالت ارائه کرده‌اند. افلاطون آن را هماهنگی و تعادل، ارسطو تناسب و شایستگی، کانت احترام به خودمختاری و جهان‌شمولی، و رالز انصاف و حمایت از محرومان می‌داند (Bunnin & Yu, 2004, pp. 367; Pomerleau, 2025). برخی پژوهشگران معانی دیگر نیز برای عدالت ذکر کرده‌اند (Barry & Matravers, 2005, pp. 481-486). این اختلاف نظرها، اجرای عدالت در هوش مصنوعی را پیچیده می‌کند (Coeckelbergh, 2021, 43-47).

یک نظرسنجی نشان می‌دهد که هیچ یک از مکاتب اصلی اخلاقی از حمایت قاطع اکثریت برخوردار نیستند. برای مثال، حدود یک چهارم فیلسوفان وظیفه‌گرایی یا نتیجه‌گرایی[2] را «می‌پذیرند» یا به آن «تمایل دارند». حدود یک سوم آنها اخلاقِ فضیلت[3] را می‌پذیرند یا به آن تمایل دارند (Bourget & Chalmers, 2014). این اختلاف نظرها تعریف «عدالت» یا «اخلاق» را برای ماشین‌ها دشوار می‌کند؛ زیرا هیچ توافق جامعی دربارهٔ نظریه اخلاقی وجود ندارد.

بر خلاف هوش مصنوعی، انسان‌ها با «شهود اخلاقی»[4] به دنیا می‌آیند و در طول سالیان زندگی چیزهایی دربارهٔ درست و نادرست می‌آموزند. انسان‌ها به کمک شهود اخلاقی- که نوعی حس فطری است- در موقعیت‌های مختلف، تصمیم‌های اخلاقی می‌گیرند. برخی بر این باورند که شهود اخلاقی انسان مانند «جعبه سیاه» پیچیده و تا حدی «غیرقابل فهم»[5] است. با وجود عدم درک کامل از چگونگی تصمیم‌گیری اخلاقی انسان‌ها و ذخیره‌سازی و پردازش اطلاعات توسط مغز، در عمل توافق نسبی بر سر درستی و یا نادرستی بسیاری از اعمال وجود دارد. در حالی که مناقشات اخلاقی در باب مسائلی چون سقطِ جنین، بِه‌مرگی و مانند آن در جریان است، مسائل اخلاقی دیگری نیز هستند که اختلاف‌نظر کمتری دربارهٔ آنها وجود دارد.

اگر دامنهٔ کاربرد هوش مصنوعی محدود شود و اطلاعات تصمیم‌گیری در دسترس باشد، می‌توان انتظار داشت که هوش مصنوعی تصمیم‌های اخلاقی نسبتاً دقیقی بگیرد. در بسیاری موارد، قوانین روشنی وجود دارند که می‌توان از آنها به عنوان معیارهای هنجاری استفاده کرد. هوش مصنوعی تا کنون در حوزه‌هایی مانند اعتبارسنجی مالی، دادگاه‌ها، استخدام و خدمات بهداشتی به کار رفته و پیامدهای

1. justice
2. consequentialism
3. virtue ethics
4. moral intuition
5. Inscrutable biological code

اخلاقی مهمی داشته است. این پیامدها عمدتاً ناشی از سوگیری در سامانه‌هاست. سوگیری سبب می‌شود هوش مصنوعی در قالب یادگیری ماشینی، تعصب را ایجاد، حفظ و تشدید کند و در نتیجه، افراد یا گروه‌های خاصی را با تبعیض مواجه سازد. سوگیری می‌تواند به روش‌های مختلفی بروز کند: در داده‌های آموزشی، در الگوریتم، در داده‌هایی که الگوریتم روی آنها اعمال می‌شود، و در گروه‌هایی که فناوری را برنامه‌ریزی می‌کنند (Coeckelbergh, 2021, 38) در اینجا تنها به دو مورد اشاره می‌کنیم.

### ۳. ۴. ۱. چالش‌های بهره‌گیری از هوش مصنوعی در اعتبارسنجی مالی

استفاده از سامانه‌های هوش مصنوعی برای اعتبارسنجی مالی در حال گسترش است. بانک‌ها و مؤسسات اعتباری در برخی کشورها از این سامانه‌ها برای ارزیابی درخواست‌های اعتباری بر اساس داده‌های موجود دربارهٔ درخواست‌کننده استفاده می‌کنند.

این رویکرد، مزایایی مانند کاهش مخاطرات مالی و توانایی تصمیم‌گیری سریع‌تر و مبتنی بر اطلاعات بیشتر دارد که آن را معقول‌تر می‌سازد. با این حال، احتمال بروز سوگیری‌های نظام‌مند از معایب این روش است. برای مثال، اطلاعات شخصی متقاضی اعتبار، اغلب شامل اطلاعاتی دربارهٔ محل سکونت آنهاست. بر این اساس، و با استفاده از داده‌های عمومی بیشتر یا داده‌های خصوصی جمع‌آوری‌شده، ممکن است سوگیری‌های نظام‌مندی علیه افراد ساکن در مناطق مسکونی خاص رخ دهد (Bartneck et al, 2021, 34)

بررسی ژرف سوگیری نژادی مرتبط با استفاده از الگوریتم‌ها[1] برای مدیریت مراقبت‌های پرخطر، بسیاری از چالش‌ها و مسائل مرتبط با مفاهیم ترکیبی عدالت، انصاف الگوریتمی[2] و دقت را آشکار می‌کند (Obermeyer et al. 2019). پژوهشگران الگوریتمی را که برای شناسایی بیماران پرخطرِ نیازمندِ مراقبت فوری استفاده می‌شد و بر درمان میلیون‌ها آمریکایی تأثیر می‌گذاشت، بررسی کردند. اگر چه این الگوریتم به‌طور مشخص نژاد را در محاسبات خود لحاظ نمی‌کرد، اما از اطلاعات مربوط به هزینه‌های مراقبت‌های بهداشتی برای پیش‌بینی نیاز به این مراقبت‌ها استفاده می‌کرد. سامانه از یادگیری ماشینی برای ایجاد الگویی به منظور پیش‌بینی هزینه‌های آینده مراقبت‌های بهداشتی استفاده می‌کرد. فرض سامانه این بود که افرادی با بیشترین هزینه‌های مراقبت بهداشتی، بیشترین نیاز را به این مراقبت‌ها دارند. این فرض منطقی به نظر می‌رسد، اما در عمل سبب نابرابری‌هایی می‌شود که با نژاد

1. algorithms
2. algorithmic fairness

همبستگی پیدا می‌کند. برای مثال، بیماران کم‌درآمد در دسترسی به مراقبت‌های بهداشتی با چالش‌های بیشتری روبه‌رو هستند؛ زیرا ممکن است به وسایل حمل و نقل مناسب یا مراقبت از کودکان دسترسی نداشته باشند یا با محدودیت‌های ناشی از مشاغل سخت و رقابتی مواجه باشند. پژوهشگران نتیجه می‌گیرند که مسئله اصلی، صورت‌بندی معضل است: چالش توسعه الگوریتم‌های محاسباتی دقیق که بر اساس مفاهیم مبهم و غیرشفاف عمل می‌کنند. ناگزیر انواع معیارهای دقیقی که برای چنین الگوریتم‌هایی نیاز است، شامل تحریف‌های است که اغلب نابرابری‌های ساختاری و عوامل مرتبط با آن را بازتاب می‌دهد. این مسائل ممکن است در میان بسیاری از الگوریتم‌های صنعتی در صنایع مختلف یافت شوند (Bartneck et al, 2021, 34).

این مطالعه نشان می‌دهد که چالش اصلی در طراحی الگوریتم‌های عادلانه، غلبه بر سوگیری‌های ناخودآگاه است. برای رفع این چالش، اقدامات زیر پیشنهاد می‌شود:

الف) داده‌های آموزشی باید متنوع و نماینده جمعیت‌های مختلف باشد و از سوگیری پاک شوند.

ب) عملکرد الگوریتم‌ها باید به‌طور مداوم بررسی و در صورت مشاهدهٔ سوگیری، اصلاح شوند.

ج) در پرتو شفافیت، کاربران باید بتوانند درک کنند که الگوریتم‌ها چگونه تصمیم‌گیری می‌کنند.

د) نهادهای نظارتی باید بر عملکرد الگوریتم‌ها نظارت داشته، از رعایت اصول اخلاقی و قانونی مطمئن شوند.

این مسائل نشان می‌دهند که طراحی الگوریتم‌های عادلانه، چالش پیچیده‌ای است که نیازمند همکاری متخصصان حوزه‌های مختلف، از جمله علوم رایانه، آمار، علوم اجتماعی و حقوق است.

### ۳. ۴. ۲. چالش‌های اخلاقی استفاده از هوش مصنوعی در نظام قضایی

در برخی کشورها، استفاده از نرم‌افزارهای هوش مصنوعی در نظام قضایی و دادگاه‌ها در حال گسترش است. یکی از کاربردهای نسبتاً ساده این نرم‌افزارها، تعیین اولویت رسیدگی به پرونده‌ها توسط قضات است. این سامانه‌ها با تحلیل اطلاعاتی مانند شدّت جرم، سوابق کیفری متهم و دیگر عوامل مرتبط، به قضات کمک می‌کنند تا پرونده‌ها را به‌طور بهینه اولویت‌بندی کنند. هدف این فناوری افزایش کارایی دادگاه‌ها و کاهش زمان رسیدگی به پرونده‌ها است (Lin et al., 2019). کاربرد پیشرفته‌تر این فناوری، کمک به قضات در تصمیم‌گیری دربارهٔ آزادی مشروط زندانیان است. در این موارد، سامانه‌های هوش مصنوعی با تحلیل داده‌هایی مانند سابقهٔ کیفری، رفتار زندانی در دوران حبس، و دیگر عوامل

مرتبط، احتمال بازگشت به جُرم را ارزیابی می‌کنند. با این حال، استفاده از این فناوری همواره با چالش‌هایی همراه بوده است. برای مثال، مطالعه‌ای توسط پروپابلیکا[1] در سال ۲۰۱۶ نشان داد که سامانهٔ کُمپاس[2] که برای ارزیابی خطر بازگشت به جرم[3] استفاده می‌شود، سوگیری نظام‌مندی علیه متهمان آمریکایی-آفریقایی در شهرستان براوِرد[4] ایالت فلوریدا داشته است (Angwin et al. 2016; Coeckelbergh, 2021, pp. 38). این سامانه خطر بازگشت به جُرم را برای این گروه بالاتر ارزیابی می‌کرد؛ در حالی که برای متهمان سفیدپوست با سوابق مشابه، خطر کم‌تری پیش‌بینی می‌شد. این سوگیری می‌تواند به محرومیت بیشتر متهمان سیاه‌پوست از آزادی مشروط یا احکام سنگین‌تر شود.

این پرونده بحث‌های گسترده‌ای را در محافل علمی، حقوقی و اجتماعی برانگیخت. توسعه‌دهندگان «کُمپاس» در پاسخ به تحلیل آماری پروپابلیکا استدلال کردند که این نهاد تنها «بر آمارهای دسته‌بندی تمرکز کرده که نرخ‌های پایهٔ متفاوت تکرار جُرم برای سیاه‌پوستان و سفیدپوستان را درنظر نگرفته است» (Dieterich et al, 2016, pp. 1). این استدلالی بسیار فنی و مبتنی بر تفاوت‌های آماری بین گروه‌های مختلف بود. برخی مفسّران اشاره کرده‌اند که در یادگیری ماشینی تعریف‌های متعددی از «انصاف» وجود دارد. برخی تعاریف بر برابری در دقت پیش‌بینی تأکید می‌کنند، در حالیکه برخی دیگر بر برابری در نرخ‌های مثبت و منفی کاذب تمرکز دارند. با توجه به نرخ‌های تکرار جُرم در شهرستان براوِرد، از نظر ریاضی ثابت شده ابزاری که برابری پیش‌بینی را برآورده می‌کند، نمی‌تواند نرخ‌های مثبت و منفی کاذب[5] برابر در گروه‌هایی با شیوع تکرار جُرم

---

۱. پروپابلیکا (ProPublica) اتاقِ خبر مستقل و غیرانتفاعی است که در سال ۲۰۰۷ با تمرکز بر روزنامه‌نگاری تحقیقی تأسیس شد. مأموریت این سازمان بررسیِ عمیق موضوعات حیاتی مانند سیاست، محیط زیست، عدالت کیفری و بهداشت، و افشای سوءاستفاده از قدرت است. پروپابلیکا با بیش از ۱۵۰ کارمند تحریریه، گزارش‌هایی تولید می‌کند که منجر به تصویب قوانین جدید، تغییر سیاست‌ها و پاسخ‌گویی نهادهای قدرت شده است. این سازمان با همکاری بیش از ۹۰ رسانه، داستان‌هایی با تأثیر ملموس بر جامعه را پوشش می‌دهد و موفق به دریافت جایزه پولیتزر، معتبرترین جایزهٔ روزنامه‌نگاری، شده است.

۲. کمپاس (COMPAS) مخفف «پروفایل مدیریت مجرمان اصلاحی برای مجازات‌های جایگزین» نرم‌افزار مدیریت پرونده و تصمیم‌گیری است که توسط شرکت نورث پوینت (که اکنون با عنوان Equivant شناخته می‌شود) تولید شد. دادگاه‌های آمریکا، از جمله نیویورک، ویسکانسین، کالیفرنیا و شهرستان براورد فلوریدا از آن برای ارزیابی احتمال تکرار جرم استفاده کرده‌اند.

3. recidivism

4. Broward County

۵. «نرخ‌های مثبت کاذب و منفی کاذب» (false positive and negative rates) به خطاهایی در الگوهای پیش‌بینی مانند سامانه کمپاس اشاره دارد. «نرخ مثبت کاذب» میزان افرادی است که مجدداً مرتکب جرم نمی‌شوند اما به‌اشتباه پرخطر شناخته می‌شوند. «نرخ منفی کاذب» میزان افرادی است که مجدداً مرتکب جرم می‌شوند اما به اشتباه کم‌خطر شناخته می‌شوند. بحث اصلی این است که این نرخ‌ها می‌توانند بین گروه‌های جمعیتی مختلف متفاوت باشند. اگر نرخ‌های پایه تکرار جرم بین گروه‌ها

متفاوت داشته باشد (Chouldechova, 2017, pp. 154). به عبارت دیگر، اگر نرخ پایهٔ تکرار جُرم در دو گروه نژادی متفاوت باشد، دستیابی همزمان به برابری در دقت پیش‌بینی و برابری در نرخ‌های مثبت و منفی کاذب، از نظر ریاضی غیر ممکن است.

یافتن راه‌حل‌های مؤثر برای چالش‌های عدالت و انصاف در سامانه‌های هوش مصنوعی دشوار است؛ زیرا تعاریف مختلفی از انصاف وجود دارد که گاه ناسازگارند. برای مثال، ممکن است بین انصاف و دقت بده‌بستانی اجتناب‌ناپذیر وجود داشته باشد، به‌طوری که بهبود یکی به کاهش دیگری منجر شود. برخی پژوهشگران معتقدند تلاش برای «نابیناسازی» الگوریتم‌ها نسبت به اطلاعات نامطلوب مانند نژاد یا جنسیت ممکن است زیان‌بار باشد؛ زیرا اطلاعات مهمی را از دست می‌دهد که برای درک کامل وضعیت موردنیاز هستند. به جای این رویکرد، بهتر است نحوه استفاده از یادگیری ماشینی و هوش مصنوعی را دگرگون کنیم (Kleinberg et al, 2018).

حذف سوگیری‌های ناخودآگاه از برنامه‌نویسی هوش مصنوعی لازم است، اما داده‌ها نیز باید واقعیت موجود را منعکس کنند. دستکاری داده‌ها و نادید گرفتن واقعیت‌های اجتماعی و ساختاری، مانع نتیجه‌گیری و اقدامات مؤثر می‌شود. داده‌ها باید به‌گونه‌ای جمع‌آوری و تحلیل شوند که ضمن انعکاس واقعیت‌ها، از تقویت نابرابری‌های ساختاری نیز جلوگیری کنند.

آگاهی از محدودیت‌های دسته‌بندی بر مبنای الگوهای آماری بسیار مهم و حیاتی است. در حالی که بسیاری از مناقشات مربوط به پیش‌بینی تکرار جُرم در شهرستان براورد بر تفاوت میان نتایج مثبت کاذب بین سیاه‌پوستان و سفیدپوستان متمرکز بود، مطالعه‌ای نشان داد که دقت پیش‌بینی‌های خودِ سامانهٔ کمپاس تنها حدود ۶۵ ٪ است. به باور نویسندگان، «این پیش‌بینی‌ها آن‌طور که انتظار داریم دقیق نیستند، به‌ویژه از دیدگاه متهمی که سرنوشتش به این پیش‌بینی‌ها وابسته است» (Dressel & Farid 2018, pp. 3). این یافته‌ها نشان می‌دهند که سامانه‌های پیش‌بینی مبتنی بر الگوهای آماری، حتی با طراحی ظاهراً بی‌طرفانه، ممکن است به دلیل خطاهای نظام‌مند یا سوگیری‌های پنهان در داده‌ها، نتایج ناعادلانه‌ای تولید کنند. برای مثال، اگر سامانه کُمپاس به اشتباه فردی را «خطر بالا» شناسایی کند، می‌تواند منجر به محرومیت از آزادی مشروط یا مجازات سنگین‌تر شود. چنین پیش‌بینی اشتباهی می‌تواند تأثیرات جدّی بر زندگی افراد و خانواده‌های‌شان داشته باشد.

---

متفاوت باشد، دستیابی به برابری پیش‌بینی بدون تأثیر بر نرخ‌های مثبت و منفی کاذب ممکن است امکان‌پذیر نباشد. این امر اطمینان از انصاف در پیش‌بینی‌ها را پیچیده می‌کند، زیرا شیوع تکرار جرم در بین همه گروه‌ها برابر نیست و می‌تواند منجر به نابرابری در دقت طبقه‌بندی افراد توسط مدل شود.

این مطالعه نشان می‌دهد که استفادهٔ صرف از الگوریتم‌ها در تصمیم‌گیری‌های مهم، مانند قضایی، خطرناک است؛ زیرا نمی‌توانند پیچیدگی رفتار انسانی را کاملاً درک کنند و خطا دارند. بنابراین، در تصمیم‌گیری‌های حسّاس مانند آزادی مشروط باید افزون بر الگوریتم‌ها، به قضاوت و تجربهٔ انسانی نیز توجه شود. همچنین، این پژوهش بر شفافیت عملکرد و محدودیت‌های الگوریتم‌ها و ضرورت آگاهی کاربران از نحوه محاسبات و دقت پیش‌بینی‌ها تأکید دارد. افزون بر این، لزوم رعایت ملاحظات اخلاقی در کاربرد هوش مصنوعی برای تصمیم‌گیری‌های حسّاس و استفاده محتاطانه از الگوریتم‌ها برای جلوگیری از سوءاستفاده یا عواقب ناخواسته را یادآور می‌شود.

## ۳. ۵. از ضرورت توضیح‌پذیری تا چالش‌های شفافیت

در اینکه آیا شفافیت[1] مستلزم توضیح‌پذیری[2] است یا نه، میان صاحب‌نظران اختلاف‌نظر است (Esposito, 2022). به نظر می‌رسد این دو مفهوم در حوزهٔ هوش مصنوعی مرتبط، اما متمایز هستند. درک تفاوت بین آنها برای طراحی و پیاده‌سازی سامانه‌های هوش مصنوعی مسئولانه و قابلِ‌اعتماد انکارناپذیر است. شفافیت در هوش مصنوعی به معنای دسترسی کامل به اطلاعات مربوط به طراحی، ساختار و عملکرد سامانه است. این شامل جزئیات فنی مانند کدها، الگوریتم‌ها و داده‌هاست که متخصصان فنی و توسعه‌دهندگان اجازه می‌دهد سوگیری‌ها، خطاها یا پیامدهای ناخواسته را شناسایی و اصلاح کنند (Manure et al, 2023, pp. 12, 61). برای مثال، کدهای یک الگوریتم یادگیری ماشینی به‌گونه‌ای کامل منتشر می‌شود که هر کسی بتواند آن را بررسی و بازبینی کند.

توضیح‌پذیری به توانایی سامانهٔ هوش مصنوعی در توضیح چرایی و چگونگی رسیدن به تصمیم یا نتیجهٔ خاص به شکلی قابل فهم برای کاربران اشاره دارد. این امر به کاربران کمک می‌کند تا بفهمند چرا سامانه به نتیجه خاصی رسیده و آیا این نتیجه قابل اعتماد است یا نه. هدف از توضیح‌پذیری این است که انسان بتواند علل و عوامل پشت تصمیم‌ها یا پیش‌بینی‌های سامانه هوشمند را درک کند (Manure et al, 2023, pp. 13, 62). برای مثال، سامانهٔ تشخیص پزشکی می‌تواند با اشاره به علائم و داده‌های کلیدی، دلیل تشخیص بیماری خاص را توضیح دهد. این دو مفهوم ـ که مکمل یکدیگرند ـ برای ایجاد سامانه‌های هوش مصنوعی قابل اعتماد و مسئولانه ضروری هستند، اما اهداف و مخاطبان متفاوتی دارند. ترکیب این دو مفهوم می‌تواند به توسعهٔ

1. transparency
2. Explicability

سامانه‌هایی منجر شود که هم دقیق و هم اخلاقی باشند.

دیدگاه صاحب‌نظران دربارهٔ توضیح‌پذیری و شفافیت در سامانه‌های هوش مصنوعی متفاوت است؛ اگر چه اکثر پژوهشگران بر امکان توضیح‌پذیری توافق دارند، برخی دربارهٔ امکان شفافیت کامل تردید دارند. شفاف‌سازی کامل میلیون‌ها کد پیچیدهٔ هوش مصنوعی هم سودمند نیست و هم چالش‌برانگیز است؛ زیرا اولاً، حجم عظیم کُدها حتی برای متخصصان قابل درک نیست و نیاز به تخصص فنی بالا و زمان زیاد دارد. دوم، افشای کامل کدها می‌تواند خطر کپی‌برداری توسط رقبا را افزایش دهد و انگیزه نوآوری را کاهش دهد. قوانین مالکیت فکری نیز از اسرار تجاری شرکت‌ها حمایت می‌کند (Coeckelbergh, 2020, pp. 121). به دلیل این ملاحظات، مفهوم «توضیح‌پذیری» به عنوان راهکاری مکمل یا جایگزین برای شفافیت کامل کد مطرح شده است.

برخی توضیح‌پذیری را هم به معنای فهم‌پذیری[1] و هم پاسخ‌گویی[2] گرفته‌اند (Floridi et al, 2018, pp. 699-700). در کاربردهای اخلاقی، مهم است که کاربران و افراد متأثر از سامانه‌های هوش مصنوعی بتوانند چگونگی تصمیم‌گیری آنها را دقیقاً بفهمند. «فهم‌پذیری» به این معناست که کارکردهای هوش مصنوعی برای انسان قابل‌فهم باشد. این سامانه‌ها نباید مانند «جعبه سیاه» اسرارآمیز باشند، بلکه باید طوری طراحی شوند که حتی برنامه‌نویسی باتجربه بتواند کارکرد آنها را بفهمید و برای دیگران، از جمله قضات، هیئت‌های منصفه و کاربران عادی توضیح دهد.

اتحادیه اروپا در چارچوب مقررات عمومی حفاظت از داده‌ها[3] «حق دسترسی به اطلاعات»[4] را به اجرا گذاشت. این قانون به افرادی که منافع‌شان از تصمیم الگوریتمی[5] متأثر شده، حق می‌دهد چگونگی تصمیم‌گیری الگوریتم دربارهٔ خود را درخواست و توضیحاتی را دریافت کنند. هدف این قانون افزایش شفافیت و پاسخ‌گویی سامانه‌های هوش مصنوعی و یادگیری ماشینی و اطمینان از عادلانه و بدون تبعیض بودن تصمیم‌های خودکار است (Harasimiuk & Braun, 2021, pp. 34-36).

با این حال، اجرای این قانون برای روش‌های «غیرقابل فهم» و پیچیدهٔ یادگیری ماشینی مانند

---

1. intelligibility
2. accountability
۳. مقررات عمومی حفاظت از داده‌ها (The General Data Protection Regulation) که به اختصار GDPR نامیده می‌شود، مقررات اتحادیه اروپا در باب حریم خصوصی اطلاعات در اتحادیه اروپا (EU) و منطقه اقتصادی اروپا (EEA) است. این مقررات که در سال ۲۰۱۶ منتشر و ۲۰۱۸ اجرایی شد، جزو مهم قانون حریم خصوصی اتحادیه اروپا و قانون حقوق بشر، بویژه ماده ۸ (۱) منشور حقوق اساسی اتحادیه اروپا است. این مقررات همچنین بر انتقال داده‌های شخصی به خارج از اتحادیه اروپا و منطقه اقتصادی اروپا نظارت می‌کند (جهت آگاهی بیشتر، نک: https://gdpr-info.eu).
4. right to information
5. algorithmic decision

شبکه‌های عصبی چالش‌برانگیز است. برخی معتقدند «غیر قابل فهم بودن» این سامانه‌ها اهمیتی ندارد و می‌توان آنها را به‌روش تجربی آزمایش و ارزیابی کرد. از این دیدگاه، تا زمانی که سامانه درست کار کند و نتایج دلخواه را ارائه دهد، نیازی به توضیح دقیق چگونگی عملکرد آن نیست (Weinberger, 2018). این دیدگاه بیشتر بر کارایی عملی سامانه‌ها تأکید دارد تا بر شفافیت یا توضیح‌پذیری آنها.

این فهم‌ناپذیری در بسیاری از کاربردهای یادگیری ماشینی مشکلی ندارد، اما در همهٔ شرایط، مناسب نیست و می‌تواند مشکلاتی بیافریند:

الف) کاهش اعتماد عمومی به هوش مصنوعی به دلیل ناتوانی در توضیح فرایند تصمیم‌گیری.

ب) مشکلات اخلاقی و حقوقی در حوزه‌های پزشکی و قضایی، جایی که تصمیم‌ها تأثیرات مهمی بر زندگی افراد دارند.

در موقعیت‌های دارای بار اخلاقی و حقوقی، توضیح و توجیه[1] برای تصمیم‌ها، ضرورتی انکارناپذیر است. عملکرد اخلاقی تنها انجام کارِ درست نیست، بلکه توجیه درستی کارها نیز مهم است. توجیه اخلاقی نمی‌تواند بر اساس جعبه سیاه «غیرقابل فهم» باشد. تحقیقات دربارهٔ «هوش مصنوعی توضیح‌پذیر»[2] در حال انجام است تا بتواند چگونگی تصمیم‌گیری شبکه‌های عصبی را روشن سازد (Bartneck et al, 2021, pp. 36). ممکن است این پژوهش‌ها ـ در نهایت ـ به یادگیری ماشینی امکان دهد تا توضیحات کافی و مناسبی برای تصمیم‌های خود ارائه کند.

با این حال، در عمل، سامانهٔ کمپاس برای ارزیابی خطر تکرار جُرم و در نتیجه، تأثیرگذاری بر چشم‌انداز آزادی مشروط[3] در دادگاه‌ها استفاده شده است.[4] در پروندهٔ «لومیس علیه ویسکانسین»، شاکی استدلال کرد که به دلیل ماهیت اختصاصی[5] الگوریتم کمپاس، از «روند دادرسی عادلانه»[1]

1. justification
2. explainable AI
3. prospects for probation

۴. «چشم‌اندازهای آزادی مشروط» به احتمال اعطای آزادی مشروط به متهمان اشاره دارد. سامانهٔ کمپاس برای ارزیابی خطر تکرار جرم استفاده می‌شود و بر تصمیم قاضی تأثیر می‌گذارد. امتیازات خطر این سامانه می‌تواند بر اعطای آزادی مشروط، شدت نظارت و شرایط آن تأثیر بگذارد. امتیاز خطر بالاتر ممکن است احتمال آزادی مشروط را کاهش دهد، در حالی که امتیاز پایین‌تر می‌تواند آن را افزایش دهد. این نشان می‌دهد که چگونه ارزیابی خطر بر تصمیم‌های قضایی درباره آزادی مشروط تأثیر می‌گذارد.

۵. مراد از «ماهیت اختصاصی» الگوریتم کمپاس، عدم دسترسی عمومی به جزئیات آن است. این الگوریتم به عنوان یک راز تجاری توسط شرکتی خاص محافظت می‌شود و نحوه ارزیابی خطر، محاسبه امتیازات و نتایج آن برای عموم شفاف نیست. این فقدان شفافیت نگرانی‌هایی را در مورد عدالت و رعایت اصول دادرسی عادلانه در فرآیندهای قانونی ایجاد می‌کند.

محروم شده است. وکیل مدافع او نمی‌توانست مبنای علمی محاسبهٔ امتیاز او را به چالش بکشد. با این همه، درخواست‌های تجدیدنظر او رد شد. قضات رأی دادند که تصمیم‌های مربوط به صدور احکام صرفاً بر اساس امتیازهای خطرِ محاسبه شده توسط کمپاس اتخاذ نمی‌شوند، بلکه می‌توانند این امتیازها را در کنار عوامل دیگر در ارزیابی خود از خطر تکرار جُرم در نظر بگیرند (Bartneck et al, 2021, pp. 36).

پاسخ‌گویی می‌تواند به صورت ثبت وقایع باشد.[۲] مثال آن دستگاه‌های ضبط کنندهٔ پرواز یا «جعبه سیاه» در هواپیماهاست که اطلاعات حیاتی مانند سرعت، ارتفاع، مکالمات خلبان و غیره را ثبت می‌کنند. در صورت سانحه، این داده‌ها برای تعیین علت اصلی و مسئولیت‌ها تحلیل می‌شوند. این فرایند همچنین اطلاعات مهمی برای پیشگیری از حوادث مشابه در آینده فراهم می‌کند.

ثبت وقایع در سامانه‌های هوش مصنوعی برای افزایش شفافیت و پاسخ‌گویی اجتناب‌ناپذیر است. این فرایند به شناسایی خطاها، درک بهتر تصمیم‌گیری‌ها و افزایش اعتماد کمک می‌کند. همچنین ثبت وقایع در تعیین مسئولیت حوادث و منشأ خطا مؤثر است، اما «ثبت وقایع» به تنهایی کافی نیست و نیازمند تحلیل دقیق و ارائه شفاف نتایج به ذی‌نفعان است. افزون بر این، ساز و کارهای رسیدگی به شکایات نیز باید وجود داشته باشد. در مجموع، ثبت وقایع ابزاری قدرت‌مند برای افزایش شفافیت، پاسخ‌گویی و مسئولیت‌پذیری در هوش مصنوعی است.

## ۴. نتیجه

از مطالب مطرح‌شده می‌توان دریافت که کاربران تنها زمانی به سامانه‌ها اعتماد می‌کنند که آنها را سودمند یابند و از عهدهٔ هزینه‌های آن برآیند. بنابراین، صاحبان کسب و کارها باید سامانه‌هایی طراحی کنند که مورد اعتماد مردم باشند. اعتماد مفهومی پیچیده است و برای ایجاد آن، رعایت عوامل مختلف ضروری است. سامانه‌ای اعتمادپذیر از نظر کاربران باید سودمند باشد، به زندگی یا

---

۱. پروندهٔ لومیس علیه ویسکانسین پروندهٔ حقوقی بود که در سال ۲۰۱۶ توسط دیوان عالی ایالت ویسکانسین بررسی شد. اریک لومیس (Eric Loomis) به اتهام رانندگی خودرویی مرتبط با حادثه تیراندازی دستگیر شد و به فرار از دست پلیس و رانندگی بدون اجازه مالک اعتراف کرد. لومیس استدلال کرد که حق او برای رسیدگی عادلانه به دلیل ماهیت اختصاصی الگوریتم کمپاس نقض شده است. او مدعی شد این امر مانع شده که اعتبار علمی و صحت امتیاز ارزیابی خطر را به چالش بکشد.

۲. ثبت وقایع یا رویدادنگاشت (log file) به پروندهٔ رایانه‌ای گفته می‌شود که رویدادها، فرآیندها و پیام‌ها را از برنامه‌ها، دستگاه‌های‌عامل مختلف ضبط می‌کند. این فایل‌ها برای نظارت بر عملکرد سامانه، عیب‌یابی مشکلات و تضمین امنیت با ارائه گزارشی تاریخی از فعالیت‌های درون سامانه‌ای بسیار مهم هستند.

کارشان کمک کند و مطمئن باشند که بـه آنهـا آسـیب نمی‌رسـاند یـا منافع‌شـان را از طریـق نقـض حریم خصوصی تهدید نمی‌کند.

وانگهی، کاربران باید مطمئن شوند که هـوش مصـنوعی خودمختـاری آنهـا را تهدیـد نمی‌کنـد. هوش مصنوعی باید تنها با دلایل موجه، مانند تضاد با اصـول اخلاقـی، قـوانین یـا منـافع بلندمـدت کاربران، از اجرای دستورات خودداری کند. رباتی هوشمند باید از اجرای دسـتورات غیـر اخلاقـی و غیرقانونی پرهیز کند. این ویژگی در کاربردهای حسّـاس ماننـد مراقبت‌هـای بهداشـتی، سـامانه‌های قضایی و خودروهای خودران اهمیت دارد، جایی که هوش مصـنوعی بایـد تصـمیم‌هایی همسـو بـا ارزش‌های اخلاقی و حقوق انسانی بگیرد.

مردم باید از عملکرد عادلانه سامانه‌های هوش مصنوعی اطمینـان یابنـد. تعیـین چارچوب‌هـای عملکردی برای هوش مصنوعی به دلیل نبود توافق بر سر یک نظریه اخلاقی جامع دشوار است. بـا ایـن حال، کاربردهای عملی هوش مصنوعی در حوزه‌های مختلف نشان می‌دهد که ایـن فنـاوری می‌توانـد نقش سازنده‌ای در بهبود زندگی انسان‌ها ایفا کند. توسعه و استفاده از هوش مصنوعی بایـد بـا توجـه بـه مسائل اخلاقی و اجتماعی انجام شود. برای اعتماد به هوش مصنوعی، آنها بایـد توضـیح‌پذیر باشـند و سوابق تصمیم‌گیری‌های خود را نگهداری کنند تا بتوانند توضیحات روشن ارائه دهند.

## فهرست منابع

کانت، اِمانوئل. (۱۳۹۴)، بنیاد مابعدالطبیعه اخلاق. (ترجمه: حمید عنایت و علی قیصری، چاپ دوم). تهران: انتشارات خوارزمی.

سالیوان، راجر. (۱۳۸۰). اخلاق در فلسفه کانت. (ترجمه: عزت‌اللّه فولادوند). تهران: طرح نو.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *Pro Publica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barry, B., & Matravers, M. (2005). Justice. In E. Craig (Ed.), *The Shorter Routledge Encyclopedia of Philosophy* (pp. 481–486). Routledge.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *An introduction to ethics in robotics and AI*. Springer.

Beauchamp, T. L., & Childress, J. F. (2012). *Principles of biomedical ethics* (7th ed.). Oxford University Press.

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe? *Philosophical Studies*, 170(3), 465–500.

Bunnin, N., & Yu, J. (2004). Justice. In *The Blackwell dictionary of western philosophy*. Blackwell Publishing.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163.

Christman, J. (2020, June 29). Autonomy in moral and political philosophy. In *Stanford encyclopedia of philosophy*. Retrieved December 24, 2024, from https://plato.stanford.edu/entries/autonomy-moral/

Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.

Coeckelbergh, M. (2022). *The political philosophy of AI: An introduction*. Polity.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Equivant. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580.

Esposito, E. (2022). Does explainability require transparency? *Sociologica*, 16(1), 17–27.

Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.

Formosa, P. (2021). Robot autonomy vs. human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy. *Minds and Machines*, 31(4), 595–616.

Gorenc, N. (2022). Hate speech or free speech: An ethical dilemma? *International Review of Sociology*, 32(3), 413–425.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5), 517–527.

Harasimiuk, D. E., & Braun, T. (2021). *Regulating artificial intelligence: Binary ethics and the law*. Routledge.

Haring, K. S., Mougenot, C., Ono, F., & Watanabe, K. (2014a). Cultural differences in perception and attitude towards robots. *International Journal of Affective Engineering*, 13(3), 149–157.

Haring, K. S., Silvera-Tawil, D., Matsumoto, Y., Velonaki, M., & Watanabe, K. (2014b). Perception of an android robot in Japan and Australia: A cross-cultural comparison. In *Social Robotics* (pp. 166–175). Springer.

International Covenant on Civil and Political Rights (ICCPR). (1976, March 23). United Nations Treaty Collection. https://treaties.un.org/doc/Treaties/1976/03/19760323%2006-17%20AM/Ch_IV_04.pdf

Iphofen, R., & Kritikos, M. (2019). Regulating artificial intelligence and robotics: Ethics by design in a digital society. *Contemporary Social Science*, 16(2), 170–184.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings*, 108, 22–27.

Lagioia, F., Rovatti, R., & Sartor, G. (2023). Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. *AI & SOCIETY*, 38, 459–478.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.

Lin, Z., Chohlas-Wood, A., & Goel, S. (2019). Guiding prosecutorial decisions with an interpretable statistical model. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. https://footprints.stanford.edu/papers/smart-

prosecution.pdf

Manure, A., Bengani, S., & Saravanan, S. (2023). *Introduction to responsible AI: Implement ethical AI using Python*. Apress.

Marsh, S. (2017, August 14). Half of UK girls are bullied on social media, says survey. *The Guardian*. https://www.theguardian.com/uk-news/2017/aug/14/half-uk-girls-bullied-social-media-survey

Marshall, A. (2017, October 18). To save the most lives, deploy (imperfect) self-driving cars ASAP. *Wired*. https://www.wired.com/story/self-driving-cars-rand-report/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Pomerleau, W. P. (n.d.). Western theories of justice. In *Internet encyclopedia of philosophy*. Retrieved January 8, 2025, from https://iep.utm.edu/justwest/

Reath, A. (2005). Ethical autonomy. In E. Craig (Ed.), *The Shorter Routledge Encyclopedia of Philosophy*. Routledge.

Thielmann, S. (2016, July 8). Use of police robot to kill Dallas shooting suspect believed to be first in US history. *The Guardian*. https://www.theguardian.com/technology/2016/jul/08/police-bomb-robot-explosive-killed-suspect-dallas

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Turner, J. (2019). *Robot rules: Regulating artificial intelligence*. Palgrave Macmillan.

Wang, L., Rau, P. P., Evers, V., Robinson, B. K., & Hinds, P. (2010). When in Rome: The role of culture & context in adherence to robot recommendations. In Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (pp. 359–366). IEEE Press.

Weinberger, D. (2018, October 19). Don't make AI artificially stupid in the name of transparency. *Wired*. https://www.wired.com/story/dont-make-ai-artificially-stupid-in-the-name-of-transparency/

Welsh, S. (2018). *Ethics and security automata: Ethics, emerging technologies and international affairs*. Routledge.

Woolley, S. (2005). Children of Jehovah's Witnesses and adolescent Jehovah's Witnesses: What are their rights? *Archives of Disease in Childhood*, 90(7), 715-719.