مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

عصر فضای مجازی
چهل و دوم

# نقد و بررسی اولین پیش نویس توصیه نامه اخلاق هوش مصنوعی (یونسکو)

The first draft of the Artificial Intelligence Ethics Recommendation (UNESCO)

بسم الله الرحمن الرحیم

# عصر فضای مجازی

گزارش شماره ۴۲

دی ۱۳۹۹

مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

# نقد و بررسی اولین پیش نویس توصیه نامه اخلاق هوش مصنوعی (یونسکو)

**محتوای انتشار یافته در این اثر**
**الزاماً بیانگر دیدگاه مرکز ملی فضای مجازی نیست**


**تهیه شده درپژوهشگاه فضای مجازی**
(گروه مطالعات اخلاقی)

تهیه‌کننده: بهزاد خداقلیزاده
ناظر علمي: محمدمهدی نصر هرندی
(مدیر گروه مطالعات اخلاقی)



حقوق مادي و معنوي اين اثر متعلق به مركز ملي فضاي مجازي است و استفاده از مطالب آن صرفاً با ذكر مأخذ بلامانع است.



نشانی: تهران، میدان آرژانتین، خیابان بیهقی، نبش
خیابان ۱۶غربی، پلاک ۲۰
تلفن: ۰۲۱-۸۶۱۵۱۰۶۱
کد پستی: ۱۵۱۵۶۷۴۳۱۱

# فهرست

سخن نخست ..... ۵
چکیده ..... ۹

نکات کلی ..... ۱۳

نکات بندبه‌بند ..... ۱۹

# سخن نخست

# سخن نخست

فضای مجازی با شتاب شگرف و رو به تزایدی که در حال بسط و گسترش است تمام ساحات اجتماعی، اقتصادی، سیاسی و فرهنگی زندگی بشر را درنوردیده و هر روز بخش بزرگی از زندگی واقعی را در خود فرو برده و حیات متفاوت و جدیدی به آن می‌دهد. لذا به نظر می‌رسد دو نگاه کلان به فضای مجازی وجود دارد: نگاه اول که بالاخص در ابتدای رشد و تکوین فضای مجازی مسلط شده بود، آن را همچون ابزاری کنار سایر ابزارهای بشری تصویر می‌کرد که تنها طریقیت داشت. اما نگاه دوم، در نتیجه رشد تحولات خیره‌کننده فضای مجازی و سایه گستری آن در حوزه‌ها و شئون بشر در یک دهه اخیر آن را چون سکویی می‌داند که بسیار فراتر از شأن ابزاری حیات انسان‌ها را سامان جدیدی داده و ادعای تمدن نوینی را دارد. رویکردی که از قضا از چشمان بصیر رهبر انقلاب نیز دور نمانده و انتظاری تمدنی از فضای مجازی در ایران را مطالبه داشته‌اند.

در همین راستا گزارش‌های عصرفضای مجازی تلاش می‌کند تا فهم سازمان‌ها و دستگاه‌های مرتبط با حوزه فضای مجازی را ارتقاء بخشیده و آن‌ها را برای مواجهه فعال و خردمندانه با تحولات این عرصه مهیا سازد.

سید ابوالحسن فیروزآبادی
دبیر شورای عالی و رئیس مرکز ملی فضای مجازی

# چکیده

# چکیده

پیش‌نویس توصیه‌نامه‌ی اخلاق هوش مصنوعی یونسکو، اولین تلاش مستقل این نهاد در زمینه‌ی تدوین توصیه‌نامه‌های اخلاق هوش مصنوعی است. از آن‌جا که ارزش‌هایی مانند تنوع و دربرگیرندگی فرهنگی برای این نهاد بین‌المللی بسیار مهم است، این نهاد از کشورهای مختلف خواسته است تا نقطه نظرات کارشناسی خود را در رابطه با این سند اخلاقی ارائه دهند. نوشته‌ی حاضر، حاصل مطالعات و بحث‌های چندی از کارشناسان و پژوهشگران حوزه‌ی فلسفه و اخلاق علم و فناوری گروه مطالعات اخلاقی پژوهشگاه مرکز ملی فضای مجازی درباره‌ی این سند است که در پاسخ به درخواست یونسکو انجام شده است. نکات ارائه‌شده در این سند به دو بخش کلی و بند به بند تقسیم می‌شود. در بخش کلی انتقادات، پیشنهادها و نکاتی که درباره‌ی کلیت و ساختار سند است، ارائه می‌شود. در بخش بند به بند نیز انتقادات، پیشنهادها و نکاتی درباره‌ی بندهای مبهم، موردسؤال، تکراری، نابسنده و ... بیان شده است.

# نکات کلی

## نکات کلی

• مهم‌ترین ویژگی منفی سند، تکرار بیش از حد بسیاری از موارد در سرتاسر سند است. به‌عنوان مثال نکته‌ای که در قسمت ارزش‌ها و اصول آورده شده است، در قسمت اقدامات سیاستی نیز تکرار شده است. این تکرارها علاوه بر این‌که بر ناکارآمد بودن سند می‌افزاید، باعث لوث شدن عبارات و واژه‌های مهم می‌شود و حتی می‌تواند نکات درخشانی را که در سند بدان‌ها اشاره شده است، بی‌اهمیت یا کم‌اهمیت جلوه دهد. این موضوع، با ویژگی‌های یک سند حقوقی در تضاد است.

• نسبت ارزش‌ها و اصول محترم‌شمرده‌شده در سند با هوش مصنوعی باید در متن سند روشن‌تر باشد و از کلی‌گویی‌های نابجا پرهیز کرد تا بتوان نقش دقیق هوش مصنوعی را در مسائل ارزیابی کرد. توصیه می‌شود در هر بند، بایدها و نبایدهای مرتبط با هر ارزش به صورت مصداقی بیان شود. همچنین توصیه می‌شود دسته‌بندی گزاره‌های توصیفی و گزاره‌های هنجاری هر بند، به شکلی دقیق رعایت شود. پیشنهاد می‌شود از بند ۹۱ این توصیه‌نامه اخلاقی به عنوان الگویی برای نگارش بندهای دیگر استفاده شود.

• برخورد سند با ذی‌ربطان مختلف بر اساس میزان اثرگذاری و عواقب حاصل‌شده از آن‌ها نیست. اگر قرار است این سند از حقوق هر یک از انسان‌ها صیانت کند، باید تأکید گزاره‌های خود را بر ذی‌ربطانی که اثرگذاری منفی احتمالی بیشتری دارند، شدیدتر کند و وزن ذی‌ربطان مهم‌تر را در بخش‌های مختلف سند لحاظ کند. به بیانی مصداقی‌تر نمی‌توان اثرگذاری بازیگران قدرتمند بخش خصوصی (مانند گوگل و آمازون) را با اثرگذاری جامعه‌ی مدنی یکی در نظر گرفت.

• این سند برای جامعیت بیشتر، باید در کنار بیان محکم هنجارها، راه‌های جای‌گزینی را برای زمانی که این بایدها نقض شدند، ارائه دهد. این مسئله در بندهایی که مخاطبِ بایدهایشان، ذی‌ربطان قدرتمندتر هستند، اهمیت بسیار زیادی پیدا می‌کند. به‌عنوان مثال اگر مخاطبِ بایدهای یکی از بندها، شرکت‌های هوش مصنوعی باشد، با توجه به قدرتی که در این شرکت‌ها متمرکز شده است، باید راه‌های جای‌گزینی توسط سند برای حالت نقض هنجار یا سرپیچی از هنجار در نظر گرفته شود.

• مشخص نیست ضمانت اجرایی اصول و اقدامات سیاستی ذکرشده در سند چیست. این سند چه نسبتی با قدرت تمرکزیافته در دولت‌ها و شرکت‌ها برقرار می‌کند؟ اهرم فشار یونسکو برای اجرای اقدامات ذکر شده در سند چیست؟ با توجه به این سؤالات، پیشنهاد می‌شود که سازوکارهای اجرایی‌سازی این سند به صورت دقیق و مشخص در آن ذکر شود.

• روابط میان ذی‌ربطان مختلف - به طور خاص، ذی‌ربطان قدرتمندتر- در

سند مبهم است. مخاطب سند در بخش اقدامات سیاستی به شکل صریحی دولت‌های عضو یونسکو است؛ یعنی به‌طور مثال، مسئولیت نظارت بر شرکتی مانند گوگل و اپل با دولت ایالات متحده‌ی آمریکا است. با توجه به این موضوع، به نظر می‌رسد مواردی مانند فراملیتی بودن برخی شرکت‌ها، امکان سرپیچی آن‌ها از دستورات نظارتی دولت‌ها، امکان هم‌راستایی سیاست‌های دولت و بخش خصوصی و به طور کلی روابط و نسبت‌هایی که می‌تواند میان این ذی‌ربطان قدرتمند به وجود آید، در سند لحاظ نشده است.

• در سند به چند دسته از کشورها و توجه بیشتر بدان‌ها تأکید شده است. این کشورها عبارتند از: کشورهای با درآمد کم یا متوسط (LMICs)، کشورهای کم‌تر توسعه‌یافته (LDCs)، کشورهای محصور در خشکی در حال توسعه (LLDCs) و کشورهای جزیره‌ای در حال توسعه (SIDs). پیشنهاد می‌شود به فهرست این کشورها، کشورهای تحریم‌شده توسط قدرت‌های جهانی نیز اضافه شود؛ چرا که بسیاری از نابرابری‌ها و شکاف‌های دانشی و فناورانه میان این کشورها و کشورهای دیگر به دلیل وجود تحریم‌ها ایجاد می‌شود. با توجه به اهمیت مفهوم دربرگیرندگی برای سازمان یونسکو، لحاظ کردن این پیشنهاد امری ضروری به نظر می‌رسد.

• در سرتاسر سند بر ضرورت احترام به مجموعه‌ای از ارزش‌های جهان‌شمول[1] تأکید بسیار زیادی شده است. شاید از منظر غربی، این ارزش‌ها محترم و درست به نظر آیند؛ اما هنگامی که رویکردهای فرهنگی مختلف در نظر گرفته می‌شوند، می‌توان تزاحم میان آرا و ارزش‌ها را درک کرد، موضوعی که با تأکید بسیار زیاد بر مجموعه‌ای

1. Universal

از ارزش‌هـای جهان‌شـمول در تضـاد اسـت. ایـن تأکیـد بسـیار زیـاد می‌توانـد بـا مشـی تکثرگـرای یونسـکو و توجهـی کـه ایـن سـازمان بـه بافتارهـای فرهنگـی مختلـف دارد، در تضـاد باشـد. همچنیـن تکثـر ارزشـی -بـه شـرط احتـرام بـه ارزش‌هایـی حداقلـی- می‌توانـد ضمـن افزایـش شـمولیت و تنـوع در جامعـه، عقایـد مختلـف را در گفتگـوی بـا یکدیگـر قـرار دهـد و در نتیجـه، نظـر درسـت‌تر و بهتـر را بـرای همـگان نمایـان سـازد. بنابرایـن پیشـنهاد می‌شـود کـه اصـول اشاره‌شـده در ایـن سـند، بـه عنـوان اصولـی حداقلـی، بـه گونـه‌ای در نظـر گرفتـه شـوند کـه بتـوان تکثـر ارزشـی و فرهنگـی را در آن جـای داد.

• بـا توجـه بـه نقـش هـوش مصنوعـی در توسـعه‌ی سـلاح‌های خـودکار و آسـیب‌های حیاتـی کـه ایـن قسـم از فناوری‌هـا می‌تواننـد داشـته باشـند، بایـد تأکیـد ایـن توصیه‌نامـه بـه چنیـن موضوعـی بیشـتر باشـد. لازم اسـت در بخـش اقدامـات سیاسـتی بـه طـور دقیقـی ایـن مسـئله مـورد بررسـی قـرار گرفتـه و گزاره‌هـای هنجـاری مرتبـط بـا آن ارائـه شـود.

# نکات بند به بند

## نکات بندبه‌بند

• در بند ۲ مورد c، یکی از چالش‌های جدید هوش مصنوعی، جانبداری و تبعیض[1] علیه جنسیت‌ها، قومیت‌ها یا سنین خاص معرفی شده است، جانبداری و تبعیضی که برای زیست اخلاقی در عصر هوش مصنوعی باید جلوی آن گرفته شود. پیشنهاد می‌شود علاوه بر این موارد، جانبداری و تبعیض علیه گروه‌های دینی و هر یک از افراد یک جامعه نیز به عنوان چالش‌های جدید هوش مصنوعی معرفی شود.

• در بخش Aims and Objectives، جمله‌های تکراری بسیاری وجود دارد که با بندهای قبل و بعد آن‌ها همپوشانی معنایی نیز دارد.

• در بند ۹، تأکید بسیار زیادی بر لزوم احترام به ارزش‌ها و اصول توسط همه‌ی ذی‌ربطان حوزه‌ی هوش مصنوعی شده است. این تأکید اگرچه درست است، اما در یک نگاه عمل‌گرایانه، شاید با احترام چند ذی‌ربط خاص مانند شرکت‌های بزرگ خصوصی و دولت‌ها بتوان نتیجه‌ی دلخواه حاصل از احترام را به دست آورد. بنابراین، وزن تاثیرگذاری ذی‌ربطان خاص در حوزه‌ی هوش مصنوعی در این بند باید لحاظ شود.

1. Discrimination

• در بند ۱۸، مصادیق چالش‌های هوش مصنوعی برای مسائل زیست‌محیطی به طور دقیق معرفی نشده است.

• در بند ۲۰، پیشنهاد می‌شود علاوه بر فرد، حقی برای خانواده، گروه و جامعه به رسمیت شناخته شود تا مجبور نباشند حتما از فناوری‌های رایج هوش مصنوعی استفاده کنند و این اختیار را داشته باشند تا بدون دغدغه از استفاده از این فناوری‌ها سر باز زنند.

• در بندهای ۲۲ تا ۲۴، نسبتی که هوش مصنوعی با مسئله‌ی همبستگی پیدا می‌کند، مبهم است. پیشنهاد می‌شود که چالش‌های هوش مصنوعی در این رابطه به صورت دقیق‌تری ارائه شود.

• در بند ۲۶، عبارت AI Method بسیار مبهم بوده و هیچ توضیحی درباره‌ی آن ارائه نشده است و بدون هیچ مقدمه‌ای به شیوه‌های توجیه آن پرداخته می‌شود.

• در بند ۴۲ و بندهای مشابهی مانند بند ۶۸ آمده است که مسئولیت اخلاقی و قانونی باید به یکی از بازیگران انسانی یا نهادی حوزه‌ی هوش مصنوعی قابل انتساب باشد. این عبارت اگرچه می‌تواند درست باشد، اما در حال حاضر بحث مهمی درباره‌ی این موضوع وجود دارد که آیا می‌توان مسئولیت را به خود فناوری هوش مصنوعی نسبت داد یا خیر؛ به طوری که هر کدام از طرفینِ موافق و مخالف این مسئله، موضع محکمی برای دفاع از نظراتشان دارند. به نظر می‌آید با توجه مشی تکثرگرایانه یونسکو بهتر باشد که به جای موضع‌گیری به نفع یکی از طرفین دعوا، در سند ذکر شود که بحث‌های علمی حول این موضوع باید بیشتر پیگیری شود تا در آینده به جواب مستدل‌تر و عملی‌تری دست یافت.

• در بند ۴۸، پیشنهاد می‌شود به تفکیک میان مسائلی که هوش مصنوعی در آن‌ها می‌تواند مخاطرات جانی به وجود آورد و مسائلی که می‌توان با آموزش اخلاقی، احتمال وقوعشان را در جامعه کم کرد، توجه شود. هم‌چنین تأکید شود که در مسائل نوع اول، باید از بازوهای اجرایی سازمان ملل استفاده شود و در مسائل نوع دوم، آموزش اجتماعی و حوزه‌ی عمومی مد نظر قرار گیرد.

• در بند ۵۸، خودارزیابی[1] دولت و مراجع عمومی درباره‌ی سامانه‌های هوش مصنوعی کافی نیست؛ بلکه باید علاوه بر دولت، نهادهای مردمی و جوامع مدنی نیز این سامانه‌ها را ارزیابی کنند تا در مواردی که سامانه‌های دولتی هوش مصنوعی در حال ایجاد زیان عمومی هستند، این نهادها و جوامع بتوانند با ایجاد اجماع عمومی جلوی این زیان‌ها را بگیرند.

• در بند ۶۱، به لزوم مشارکت کشورهای مختلف (کشورهای با درآمد کم یا متوسط، کشورهای کم‌تر توسعه‌یافته، کشورهای محصور در خشکی در حال توسعه و کشورهای جزیره‌ای در حال توسعه) در بحث‌های مربوط به حکمرانی هوش مصنوعی اشاره شده است. پیشنهاد می‌شود به این بند نیز کشورهایی که به‌دلیل تحریم، امکان مشارکت در این بحث‌ها را ندارند، توجه شود تا شکاف دانشی و حکمرانی موجود میان این کشورها و کشورهای دیگر، افزایش نیابد.

• قسمت اول بند ۷۰ تکراری است و موجز نبودن و توضیح دادن دوباره و بیش از اندازه در این قسمت باعث شده است که بخش دیگر این بند که درباره توسعه‌ی سیاست‌های داخلی و محلی است، در متن این بند گم شود.

1. Self-assessment

• در بند ۷۱ و بندهای مشابهی، شفافیت[1] و توضیح‌پذیری[2] سامانه‌های هوش مصنوعی به عنوان تنها مسیر رسیدن به ویژگی قابل اعتماد بودن[3] این فناوری معرفی می‌شود. می‌توان علت این امر را خطاپذیری فناوری هوش مصنوعی دانست؛ یعنی از آن‌جا که هوش مصنوعی می‌تواند اطلاعات و تصمیمات اشتباهی را تولید کند، با شفافیت و توضیح‌پذیری این سامانه‌ها می‌توان قابل اعتماد بودن آن را بازیابی کرد. با وجود این، لازم است توجه شود که در ادبیات حال حاضر هوش مصنوعی، روش‌هایی مانند Generative Adversarial Networks وجود دارد که می‌تواند مسئله‌ی خطاپذیری هوش مصنوعی را حل کند. در این حالت می‌توان گفت دیگر نیازی به شفافیت در سامانه‌های هوش مصنوعی نیست و از این طریق می‌توان نزاع میان شفافیت و امنیت[4] را در فناوری هوش مصنوعی حل کرد.

• با توجه به این‌که در بند ۷۲ و به طور کلی بخش Data Policy، مسئله‌ی کیفیت یادگیری داده‌ها و تنوع در این فرآیند مورد تأکید واقع شده است، باید زمینه‌ی ورود فرهنگ‌های متنوع به این فرآیندها فراهم شود؛ مثلاً آزمایش Moral Machine - که قرار است مبنایی برای طراحی ماشین‌های خودران باشد- به زبان‌های بیشتر و متنوع‌تری انجام شود. هم‌چنین به منظور دستیابی به داده‌ها و نتایج دربرگیرنده‌تر و متنوع‌تر، پیشنهاد می‌شود در سند به ظرفیت شبکه‌های اجتماعی و سفرای اجتماعی سازمان ملل برای ایجاد مشارکت اجتماعی در این زمینه اشاره شود. استفاده از این ظرفیت باعث می‌شود که دادگان[5] نماینده‌ی طیفی عمومی‌تر - و نه نخبگانی‌تر- از مردم باشد.

1. Transparency
2. Explainability
3. Trustworthiness
4. Security
5. Dataset

• بند ۷۳ تکراری و طولانی است و محتوای آن در بندهای قبلی آمده است.

• در بند ۸۱ باید به این نکته توجه شود که به اشتراک گذاشتن اطلاعات حاکمیتی کشورها حول همکاری‌های بین‌المللی ممکن است با مسئله‌ی امنیت کشورها در تضاد باشد. هم‌چنین باید این ضمانت در نظر گرفته شود که پذیرش سیاست‌های سازمان‌های جهانی مانند بانک جهانی، منجر به بروز بحران‌های اقتصادی و اجتماعی در کشورها نشود.

• بند ۹۱ نمونه‌ی قابل قبولی است که اشکالات ساختاری و کلی‌گویی را ندارد. در این بند ابتدا به صورت شفاف و روشن درباره‌ی «بایدها» صحبت شده است. سپس درباره‌ی «نبایدها» صحبت شده است و پس از آن مصادیق چالش‌های هوش مصنوعی برای مسئله‌ی جنسیت ذکر شده است. پیشنهاد می‌شود این بند به عنوان الگویی برای نگارش بندهای دیگر در نظر گرفته شود.

• در بند ۱۱۸ به شکلی بسیار ضعیف به بحث شکل‌گیری مونوپولی و انحصار پرداخته می‌شود و سند بدون اشاره به نقش شرکت‌ها، تنها به چند مصداق غیرمربوط می‌پردازد؛ در حالی که باید صراحتاً در این بند ذکر شود که شرکت‌های بزرگ هوش مصنوعی به مونوپولی تبدیل شده‌اند. این ملاحظه‌گری سند در مواجهه با شرکت‌ها موضوعی است که در بندهای دیگر این سند نیز موج می‌زند.

• در بخش Economy and Labour به این نکته اشاره شود که پیش از ظهور هوش مصنوعی ثروت در میان افراد مختلف جامعه

توزیع شده بود و پس از ظهور هوش مصنوعی این ثروت عملاً در دست شرکت‌های خاصی متمرکز شده است. این اتفاق سبب شده است که به بسیاری از مشاغل کوچک و متوسط آسیب برسد. بنابراین پیشنهاد می‌شود که در سند بر لزوم وجود آستانه‌ای برای تمرکز ثروت در شرکت‌های هوش مصنوعی تأکید شود.

- با توجه به این که هوش مصنوعی می‌تواند تصمیم‌گیری‌هایی را ممکن سازد که در آن‌ها انسان‌ها از جزئیات تصمیمات باخبر نباشند و با توجه به این‌که این موضوع می‌تواند زمینه‌ساز ایجاد سلطه بر آن‌ها شود، پیشنهاد می شود به نقش هوش مصنوعی در ایجاد نوعی پدرسالاریِ سخت آن هم از جانب یک موجود بی‌جان بر انسان اشاره شود و به منظور جلوگیری از وقوع چنین پدیده‌ای راه‌هایی مانند آگاهی مردم از تصمیمات شرکت‌ها، دولت‌ها و سامانه‌های هوش مصنوعی در سند برجسته شوند.

مرکز ملی فضای مجازی
پژوهشگاه فضای مجازی

csri.majazi.ir

حوزه فضای مجازی به اندازه انقلاب اسلامی اهمیت دارد. این فضا مثل یک رودخانه پر از آب و خروشان است که می آید و دائماً هم بر آب آن افزوده و خروشان تر می شود. اگر ما بر این رودخانه تدبیر کنیم و برنامه داشته باشیم، زه‌کشی کنیم و هدایت کنیم این رودخانه را تا به سد بریزد، می‌شود فرصت. اگر رهایش کنیم و برنامه‌ای برای آن نداشته باشیم می‌شود یک تهدید.

csri.majazi.ir