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

علیرضا ثقه‌الاسلامی
استادیار پژوهشکده جامعه و اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران
seghatoleslami@irandoc.ac.ir

مقدمه: هدف از این مقاله، شناسایی و تدوین اصول و رهنمودهای اخلاق هوش مصنوعی است. دانش‌افزایی این پژوهش در گردآوری موضوعات و مسائل اخلاقی، تدوین اصول اخلاقی منسجم، و در نهایت مقوله‌بندی و پیشنهاد کدها و رهنمودهای اخلاقی در حوزه اخلاق هوش مصنوعی ذیل ذینفعان متعددی هم‌چون کاربران، طراحان، سیاست‌گزاران و دیگر کنشگران در این قلمرو است. برای دستیابی به این هدف، در چارچوب مطالعات کتابخانه‌ای و بررسی منابع متنوع، اسناد و دستورالعمل‌های بین‌المللی مرتبط با چالش‌های اخلاقی مطرح در هوش مصنوعی شناسایی و معرفی می‌شود. سپس، این اسناد و دستورالعمل‌ها با دو معیار روزآمد بودن و معتبر بودن سازمان‌های متولی گردآوری و مبتنی بر روش پژوهش‌های اِسنادی بررسی و تحلیل می‌شود. و در پایان، اصول و رهنمودهایی اخلاقی برای مواجهه با چالش‌های اخلاقی هوش مصنوعی تدوین و معرفی می‌گردد.

روش‌شناسی: روش پژوهش در دو گام تحقق یافته است. در گام اول این پژوهش، ابتدا در چارچوب مطالعات کتابخانه‌ای، منابع متنوع، اسناد و دستورالعمل‌های بین‌المللی مرتبط با چالش‌های اخلاقی مطرح در هوش مصنوعی شناسایی و معرفی می‌شود. این اسناد و دستورالعمل‌ها با دو معیار روزآمد بودن و معتبر بودن سازمان‌های متولی گردآوری و بررسی می‌شوند.

در گام دوم و در چارچوب بررسی‌های تحلیلی در گام پیشین، مسائل و چالش‌های اخلاقی شاخص (و غیرقابل تقلیل به‌یکدیگر) و هم‌چنین شناسایی ذینفعان متعددی هم‌چون کاربران، طراحان، سیاست‌گزاران و دیگر کنشگران در حوزه هوش مصنوعی است. در ادامه، اصول اخلاقی حداقلی و منسجمی برای ارجاع و استناد کدها و رهنمودهای اخلاق هوش مصنوعی به این اصول تهیه و تدوین می‌گردد. و در پایان، ناظر به مسائل و چالش‌های اخلاقی در هوش مصنوعی و با استناد به اصول اخلاقی مدون، کدها و رهنمودهای اخلاقی ذیل ذینفعان و کنشگران در حوزه هوش مصنوعی مقوله‌بندی و پیشنهاد می‌شود. برای ارزیابی روایی یافته‌ها و دستاورد پژوهش، اصول و رهنمودهای اخلاقی پیشنهادی برای مواجهه با چالش‌های اخلاقی در هوش مصنوعی، توسط گروهی کانونی (Focus Group) از متخصصان هوش مصنوعی، فلسفه و اخلاق تکنولوژی و علم اطلاعات نقد و

بررسی می‌شود. این بررسی‌ها تا زمان دستیابی به اِجماعی نسبی میان متخصصان گروه کانونی ادامه و فرآیند اعتباربخشی به یافته‌های پژوهش تحقق می‌یابد.

یافته‌های اصلی: در مرحله اولِ گام دوم این پژوهش، 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. به عنوان مثال، برای اصل خیرخواهی، علاوه 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 بود 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.

در مرحله دومِ گام دوم این پژوهش، رهنمودهای ویرایش، از نظر اصطلاحات یک‌دست‌سازی شد. هم‌چنین رهنمودهای تکراری حذف و رهنمودهای 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 سوم از گام دوم این پژوهش، رهنمودهای اخلاقی حاصل از مرحله دوم در گروهی 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 تکنولوژی؛ یک متخصص مهندسی تکنولوژی اطلاعات؛ یک متخصص مهندسی صنایع؛ یک متخصص سیاست‌گذاری بر پایه مدل) برگزار شد. این نشست ۴ ساعت به طول انجامید و ۶۷ رهنمود اخلاقی مرحله دوم در چارچوب پنج اصل اخلاقی، بار دیگر الویت‌بخشی، اصلاح و حذف شدند. نتیجه آن‌که اصل اخلاقی خیرخواهی با ۷ رهنمود، اصل عدم آسیب‌رسانی با ۱۲ رهنمود، اصل خودمختاری با ۷ رهنمود، اصل اخلاقی عدالت با ۱۰ رهنمود، و اصل اخلاقی توضیح‌پذیری با ۱۰ رهنمود، ویرایش و صورتبندی مجدد گردید.

بحث و نتیجه‌گیری: در پایان، دو نکته بسیار مهم، لازم به توضیح است: اول آن‌که 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 مشارکت در مباحث نقادانه خود قرار دادند. اما در صورت‌بندی نهایی رهنمودهای اخلاقی تلاش شده است، مخاطبان این رهنمودها، تمامی کنشگران و عامل‌های انسانی که با سیستم‌های هوش مصنوعی تعامل دارند را هدف قرار می‌دهد. البته، بی‌تردید نقش طراحان، متخصصان و توسعه‌دهندگان سیستم‌های هوش مصنوعی برای توجه به رهنمودهای اخلاقی پیشنهادی به‌مراتب

گسترده‌تر و مسئولانه‌تر است. نکته دوم آن‌که 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 آسیب‌رسانیِ سیستم‌های هوش مصنوعی به جامعه انسانی و محیط زیست را نشان می‌دهد. از این رو، به نظر می‌رسد سیستم‌های هوش مصنوعی در طول چرخه عمر خود، و در فرآیند توسعه و استقرارشان در سرتاسر کره زمین، پیش و بیش از هر دستاوردی، باید عدم آسیب رساندن به محیط زیست و موجودات صاحب شعور سکونت‌یافته در این سیاره را هدف نهایی خود قرار دهند.

کلیدواژه‌ها: اخلاق هوش مصنوعی؛ اصول اخلاقی؛ سند اخلاق هوش مصنوعی؛ رهنمودهای اخلاقی؛ هوش مصنوعی.

Examination 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

Introduction: The purpose of this article is to identify and formulate principles and guidelines for AI ethics. The contribution of this research lies in compiling ethical themes and issues, formulating coherent ethical principles, and ultimately categorizing and proposing ethical codes and guidelines in the field of AI ethics for multiple stakeholders such as users, designers, policymakers, and other actors in this domain. To achieve this aim, within the framework of library research, relevant international documents and guidelines addressing ethical challenges in artificial intelligence are identified and introduced. These documents and guidelines are collected based on two criteria—recency and credibility of their issuing organizations—and are examined and analyzed through document-based research methods. Finally, ethical principles and guidelines for addressing ethical challenges in artificial intelligence are formulated and presented.

Methodology: he research method was implemented in two steps. In the first step, and within the framework of a library-based study, diverse sources, documents, and international guidelines related to the ethical challenges of artificial intelligence were identified and introduced. These documents and guidelines were collected and examined based on two criteria: recency and credibility of the issuing organizations.

In the second step, and within the framework of analytical examination in the previous step, key and non-reducible ethical issues and challenges were identified, as well as multiple stakeholders—including users, designers, policymakers, and other actors in the domain of artificial intelligence. Subsequently, minimum and coherent ethical principles were prepared and formulated for serving as the basis for referencing and substantiating AI ethics codes and guidelines. Finally, with regard to ethical issues and challenges in artificial intelligence and grounded in the formulated ethical principles, ethical codes and guidelines were categorized under relevant stakeholders and actors in artificial intelligence and were proposed. To assess the validity of the findings and outcomes, the ethical principles and guidelines proposed for addressing ethical challenges in artificial intelligence were reviewed and critiqued by a focus group consisting of experts in artificial intelligence, philosophy and ethics of technology, and information science. These assessments continued until relative

consensus among the specialists in the focus group was achieved, thereby completing the validation process of the study's findings.

Main findings: In the first stage of the second step of this research, ethical guidelines were extracted from seven AI ethics documents through the search for identical and similar keywords corresponding to each of the five ethical principles. For example, for the principle of beneficence, in addition to the keyword beneficence and its derivatives and synonyms, terms such as wellbeing, dignity, and sustainability were also searched within all documents, and the guidelines containing these keywords were extracted and compiled. In total, 130 guidelines were extracted from seven AI ethics documents under the five ethical principles. These included 15 guidelines for beneficence, 49 guidelines for non-maleficence, 25 guidelines for autonomy, 28 guidelines for justice, and 28 guidelines for explicability.

In the second stage of the second step, the collected guidelines were revised and terminologically harmonized. Duplicate guidelines were removed, and guidelines with less comprehensiveness compared to others were also eliminated. Consequently, a total of 67 guidelines were retained, consisting of 8 guidelines for beneficence, 16 for non-maleficence, 13 for autonomy, 14 for justice, and 16 for explicability.

In the third stage of the second step, the ethical guidelines obtained from the second stage were discussed and reviewed in a focus group. This group included six experts in the fields of artificial intelligence and professional ethics (two experts in information science and knowledge studies; one expert in philosophy of science and technology; one expert in information technology engineering; one expert in industrial engineering; and one expert in model-based policy analysis). The session lasted 4 hours, during which the 67 guidelines of the second stage were prioritized, revised, and reduced again within the framework of the five ethical principles. The results were as follows: the principle of beneficence was finalized with 7 guidelines; non-maleficence with 12 guidelines; autonomy with 7 guidelines; justice with 10 guidelines; and explicability with 10 guidelines.

Discussion and conclusions: Two important points must be clarified at the end. First, the focus group of this study approached the guidelines from the perspective of professional ethics, emphasizing their primary audience—namely designers, specialists, and developers of AI systems—when participating in the critical discussions. However, in the final formulation of the ethical guidelines, an effort was made to target all human actors and agents who interact with AI systems. Nevertheless, the role of designers, specialists, and developers of AI systems in considering and applying these ethical guidelines is undoubtedly broader and more responsible. The second point is that, in the focus group's discussions on AI ethics, the principle of non-maleficence—similar to the initial review of the seven AI ethics documents—contained the largest number of ethical guidelines. This indicates the significance of preventing harm caused by AI systems to human

society and the environment. Accordingly, it appears that AI systems, throughout their life cycle and in all phases of their global development and deployment, must prioritize the prevention of harm to the environment and sentient beings inhabiting this planet above all other achievements.
Keywords: AI Ethics; Ethical Principles; AI Ethics Document; Ethical Guidelines; Artificial Intelligence.