

An Inquiry into the UNESCO's Moral Document of Artificial Intelligence from the Perspective of Ayatollah Misbah Yazdi's Foundations of Moral Philosophy

Hamid Mohseni  / A PhD. holder of transcendent philosophy, **IKI**. h.mohseni1297@mailfa.com

Received: 2024/05/15 - **Accepted:** 2024/11/01

Abstract

Due to the concerns over the potential threat artificial intelligence poses to human and moral values, UNESCO has recently drafted and adopted a document entitled "*Recommendation on the Ethics of Artificial Intelligence*". This document outlines certain goals, values, and norms, and calls upon member states -including Iran- to implement its provisions within their respective countries. International organizations seek to align the laws and values of societies with their own by adopting such high-level policy documents. The main issue addressed in this study is the compatibility or incompatibility of the ethical value foundations presented in the aforementioned recommendation with the moral principles of the Islamic school of thought, with a particular emphasis on the ideas of Ayatollah Misbah Yazdi. This focus is due to his comprehensive and distinctive theory on moral philosophy and the objective of ethics.

The present study first employs a *descriptive* method to explain the global ethical concerns arising from artificial intelligence technologies. It then adopts an *analytical* approach to examine various dimensions, including the distinction between weak and strong AI, the different types of AI ethics, the comparison between virtue ethics and action-based ethics, and the analysis of the source of moral values in the recommendation-based ethics. Ultimately, the study demonstrates that the Recommendation is grounded in contractualism and, in terms of both subject and objective, reflects an action-oriented ethics. However, the Islamic ethical theory falls under the realist schools of thought and, according to the virtue-oriented philosophy of some scholars, including Ayatollah Misbah Yazdi, it is teleological and perfectionist in nature. Ultimately, the proposed solution in AI ethics is to evaluate the use of such new technologies according to the criterion of drawing closer to God.

Keywords: ethics, artificial intelligence, UNESCO, ethical foundations, Ayatollah Misbah Yazdi.

نوع مقاله: پژوهشی

واکاوی سند اخلاق هوش مصنوعی یونسکو از منظر مبانی فلسفه اخلاق آیت‌الله مصباح یزدی

h.mohseni1297@mailfa.com

حمید محسنی  دکترای حکمت متعالیه مؤسسه آموزشی و پژوهشی امام خمینی *

دریافت: ۱۴۰۳/۰۲/۲۶ - پذیرش: ۱۴۰۳/۰۸/۱۱

چکیده

یونسکو به‌خاطر نگرانی از به‌خطر افتادن ارزش‌های انسانی و اخلاقی از ناحیه هوش مصنوعی، به‌تازگی سندی با عنوان «توصیه‌نامه اخلاق هوش مصنوعی» تنظیم و تصویب کرده است. در این سند برخی اهداف، ارزش‌ها و هنجارها آورده شده و از کشورهای عضو - که ایران نیز یکی از آنهاست - خواسته است مفاد این سند را در کشورهای خود اجرا کنند. نهادهای بین‌المللی سعی دارند با تصویب این‌گونه اسناد بالادستی، قوانین و ارزش‌های جوامع را با خود همراه سازند. مسئله اصلی این پژوهش سازگاری و ناسازگاری مبانی ارزش‌های اخلاقی توصیه‌نامه مذکور با مبانی اخلاقی مکتب اسلام با تأکید بر اندیشه آیت‌الله مصباح یزدی است. به‌سبب آنکه ایشان در موضوع علم اخلاق و هدف اخلاقی، نظریه اختصاصی و کاملی دارد. پژوهش حاضر نخست با روش «توصیفی» نگرانی‌های اخلاقی از ناحیه فناوری هوش مصنوعی در سطح دنیا را تبیین نموده و سپس با روش «تحلیلی» و با تفکیک بین هوش مصنوعی ضعیف و قوی، اقسام اخلاق هوش مصنوعی و مقایسه اخلاق فضیلت و اخلاق عمل‌محور و تحلیل منشأ ارزش اخلاقی توصیه‌نامه‌ای را به دست آورده و در نهایت مشخص ساخته است که این توصیه‌نامه مبتنی بر قراردادی‌گری بوده و به لحاظ موضوع و هدف، عمل‌محور است. اما نظریه اخلاقی اسلام جزو مکاتب واقع‌گرا بوده و در اندیشه برخی علمای فضیلت‌محور و از منظر آیت‌الله مصباح یزدی غایت‌گرایانه و کمال‌جوست. سرانجام راهکار ابداعی در اخلاق هوش مصنوعی این است که استفاده از این نوع ابزار نوین با معیار تقرب به خداوند ارزش‌گذاری شود.

کلیدواژه‌ها: اخلاق، هوش مصنوعی، یونسکو، مبانی اخلاق، آیت‌الله مصباح یزدی.

مقدمه

بسیاری معتقدند: در آینده نه‌چندان دور، ظهور «ابر هوش» (Superintelligence) امری اجتناب‌ناپذیر است، هرچند در زمان پیدایش این رخداد اختلاف نظر است (بادینگتون، ۲۰۱۷، ص ۲-۳). ولی این پدیده می‌تواند مخاطرات بسیاری جدی برای بشریت داشته باشد. از این رو لازم است دانشمندان ملاحظات را برای خطرات احتمالی آینده از ناحیه سامانه‌های پیشرفته هوش مصنوعی در نظر بگیرند. افزون بر این، همین هوش مصنوعی محدود («Narrow Artificial Intelligence» یا «NAI») نیز می‌تواند در زندگی بسیاری از انسان‌ها تأثیرگذار باشد (راسل و نورویگ، ۲۰۲۰).

این فناوری‌ها در کنار مزایای متعدد، خطرات و چالش‌هایی را نیز ایجاد می‌کنند. بخشی از این خطرات و چالش‌ها ممکن است ناشی از استفاده بدخواهانه از فناوری باشد؛ بخش مهم دیگری از تهدیدات نیز ممکن است از ناحیه خود فناوری هوش مصنوعی باشد. در مقاله پیش رو، این نگرانی‌ها، چالش‌ها و تهدیدها بیشتر تبیین می‌شود. از این رو «یونسکو» چارچوب‌های نظارت بین‌المللی و ملی را فراهم آورده است تا این فناوری در نهایت، به نفع بشریت تمام شود. به همین خاطر سندی را با عنوان «توصیه‌نامه اخلاق هوش مصنوعی» (Recommendation on the ethics of artificial intelligence) در چهل و یکمین جلسه گردهمایی عمومی خود (در تاریخ ۲۴ نوامبر ۲۰۲۱) تصویب کرده است (یونسکو، ۲۰۲۱).

این توصیه‌نامه یک مقدمه و ۱۴۱ بند دارد و در آن برخی اهداف و ارزش‌های اخلاقی بیان شده است. از کشورهای عضو - از جمله ایران - نیز خواسته شده است مفاد این سند را در کشورهای خود اجرا کنند. به این مطلب در بند ۱۳۵ اشاره شده است. هرچند مفاد برخی بندها اشاره دارد به اینکه در اجرای این قوانین، قوانین داخلی کشورها نیز ملاحظه می‌گردد، اما نهادهای بین‌المللی مستقیم یا غیرمستقیم می‌کوشند با تصویب این گونه اسناد بالادستی، قوانین و ارزش‌های جوامع را با خود همراه سازند. همان گونه که پیش از متن نهایی در گزارشی آمده - ولی در متن اصلی به آن تصریح نشده است - یونسکو سعی دارد رهبری اخلاقی خود را با تنظیم استانداردهایی در حوزه علم و فناوری و هوش مصنوعی با رویکردی انسان‌محوری تقویت کند.

(<https://unesdoc.unesco.org/ark:/48223/pf0000379920.41C/73-p.2>)

همچنین در بند ۴ سند آمده است: این توصیه‌نامه دولت‌های عضو را به‌عنوان کاربران هوش مصنوعی برای توسعه چارچوب‌های قانونی و نظارتی در کل چرخه هوش مصنوعی مخاطب قرار داده، به آنها رهنمود اخلاقی می‌دهد. بدین‌روی ضروری است مبانی و ارزش‌های اخلاقی این سند بین‌المللی از منظر مبانی ارزش‌های اخلاقی مکتب اسلام بررسی گردد.

در این زمینه تحقیقاتی انجام شده است. پیشینه‌ای که به‌صورت رسمی و مدون یافت شد، گزارش مختصری است با عنوان «تنقد و بررسی اولین پیش‌نویس توصیه‌نامه اخلاق هوش مصنوعی (یونسکو)» که توسط بهزاد خدالیزاده و تحت‌نظر محمدمهدی نصر هرنندی (مدیر گروه مطالعات اخلاقی) در «پژوهشگاه مرکز ملی فضای مجازی» چاپ شده

است. در این اثر برخی نقدهای جزئی به تک‌تک بندها و نیز نقدهای کلی به اصل سند مطرح شده است. اما در آن به مبانی اخلاقی این سند و سازگاری آن با نظریه اخلاقی اسلام اشاره نشده است. در این گزارش به توصیه‌نامه یونسکو ایراد گرفته شده که چرا در سراسر متن، به ارزش‌ها و اصول، روشن‌تر و جزئی‌تر پرداخته نشده است. از این رو پیشنهاد داده است بند ۹۱ به‌عنوان الگویی برای نگارش سایر بندها باشد (خداقلیزاده و دیگران، ۱۳۹۹، ص ۱۵). اما درواقع محتوای این بند با توجه به مفاد سایر بندها در جهت ترویج فرهنگ «برابری جنسیتی» است.

در مقاله مزبور، حتی این پیشنهاد مطرح شده است که در این سند، تکرر ارزشی و فرهنگی گنجانده شود (خداقلیزاده و دیگران، ۱۳۹۹، ص ۱۸). اگر منظور از پیشنهاد مزبور این است که «یونسکو» به ارزش‌های فرهنگ اسلام نیز بها بدهد، این نکته قابل پذیرش است. اما به نظر می‌رسد با این پیشنهاد به‌نوعی تکرر ارزشی در حوزه اخلاق پذیرفته شده است. پذیرش این مطلب نیز منجر به نسبی‌گرایی و کثرت‌گرایی در ارزش‌های اخلاقی می‌شود. از این رو لازمه این امر عدم امکان ارزیابی و داوری بین مکاتب مختلف اخلاقی است و همه آنها به یک اندازه معتبر خواهند بود. بنابراین مکتب اخلاقی اسلام بر سایر نظام‌های اخلاقی برتری نخواهد داشت. نهایت امر آن است که نظام اخلاقی اسلام نظامی اخلاقی خواهد بود در عرض سایر نظام‌های اخلاقی. اما آیا بنا بر نگرش دین اسلام و اندیشمندان مسلمان، حقیقت این گونه است؟

در اثر دیگری (مشکل از مجموعه مقالاتی که به‌زودی به‌طور رسمی منتشر خواهد شد) به تفصیل و مستقل این سند تحت نظر «دفتر فقه معاصر حوزه علمیه قم» مطالعه و نقد شده است (جمعی از نویسندگان، ۱۴۰۲).

اما با بررسی اجمالی آن پژوهش‌ها مشخص شد بیشتر تحقیقات مزبور با رویکرد فلسفه اخلاق و مقایسه آنها با نظریه اخلاقی مکتب اسلام به نقد مبانی اخلاقی سند مزبور نپرداخته‌اند. البته در یکی از این مقاله‌ها، نگارنده کوشیده است با رویکرد کلی و بنا بر آموزه‌های اخلاقی مکتب اسلام این سند را نقد کند (محسنی و دیگران، ۱۴۰۲). بیشتر تمرکز پژوهش مذکور در باب واقع‌گرایی و غیرواقع‌گرا بودن ارزش‌های سند بوده و به لوازم آنها نیز اشاره شده است. از یافته‌های تحقیق مزبور در پژوهش حاضر استفاده شده است. ولی بیشترین مطالب مقاله حاضر بر تفکیک بین مکتب اخلاقی وظیفه‌گرا، فضیلت‌محور و کمال‌گرایانه با تأکید بر اندیشه آیت‌الله مصباح یزدی است.

همچنین نگارنده مقاله مختصری درباره سازگاری یا ناسازگاری مبانی اخلاقی این سند با نظریه اخلاقی اسلام در همایش علمی - پژوهشی «تختین کنفرانس فضای سایبر» که توسط دانشکده مهندسی دانشکدگان فارابی دانشگاه تهران در آبان سال ۱۴۰۱ برگزار شد، ارائه کرده است (محسنی، و فولادی قلعه، ۱۴۰۱). می‌توان مقاله حاضر را به‌نوعی تفصیل و تکمیل دو مقاله مکتوب نگارنده دانست. افزون بر این، پژوهش حاضر به برخی نگرانی‌ها از ناحیه هوش مصنوعی اشاره کرده و به صورت اختصاصی با مبانی فلسفی آیت‌الله مصباح یزدی، سند مزبور را بررسی کرده است.

بنابراین نوشتار حاضر سعی دارد با روش «توصیفی - تحلیلی» این سند بین‌المللی در حوزه اخلاق هوش مصنوعی را بررسی کند. بنابراین مسئله اصلی این مقاله سنجش و واکاوی مبانی ارزشی سند اخلاق هوش مصنوعی «یونسکو» با نظریه فلسفه اخلاق آیت‌الله مصباح یزدی است.

پژوهش حاضر نخست نگرانی‌ها، اخلاق هوش مصنوعی و امکان اخلاق‌مندی ربات‌ها را تبیین کرده است. سپس مبانی ارزش‌های اخلاقی سند مذکور را تحلیل نموده و سرانجام نظریه «اخلاق هوش مصنوعی» بنا بر اندیشه این فیلسوف معاصر را ارائه داده است.

۱. نگرانی‌ها از ناحیه هوش مصنوعی

از منظر برخی ظهور «ابر هوش» و یا در رویکرد دیگری «هوش عمومی مصنوعی» (Artificial General Intelligence) امری اجتناب‌ناپذیر است. اما در اینکه چقدر طول می‌کشد تا هوش مصنوعی در همه توانایی‌ها به سطح هوشمندی انسانی برسد (هوش عمومی مصنوعی) و یا از توانایی‌های انسانی بالاتر برود (ابر هوش)، در میان اندیشمندان اختلاف نظر وجود دارد. ولی در میان برخی محققان، نگرانی‌های عمیقی از این ناحیه ناظر به نوع انسان‌ها ترسیم شده است که برخی از این نگرانی‌ها خیال‌پردازانه و برخی آینده‌پژوهانه به حساب می‌آیند (ر.ک. تگمارک، ۱۴۰۱). حتی اندیشمندان از ناحیه هوش مصنوعی محدود (Narrow AI) نیز نگرانی‌هایی را حس می‌کنند.

از این‌رو این مسئله بین اندیشمندان، فلاسفه و مهندسان موضوع بحث قرار گرفته است که آیا توسعه هوش مصنوعی را ادامه دهند؟ زیرا توسعه بیش از حد و بی‌هدف هوش مصنوعی برخی نگرانی‌های اخلاقی، حقوقی و حکمرانی در پی خواهد داشت. هرچند برخی افراد تنها به توسعه و پیشرفت جنبه مهندسی و فنی هوش مصنوعی اهمیت می‌دهند، ولی تحقیق و پژوهش در حوزه‌های دیگری همچون مباحث اخلاقی و حکمرانی نیز ضروری است؛ همان‌گونه که در سطح دنیا تحقیق و پژوهش در بُعد مباحث نظری، فلسفی، مسائل حکمرانی و اخلاقی در حوزه هوش مصنوعی از اهمیت فراوانی برخوردار است.

برخی رهبران فناوری و کارشناسان مطرح هوش مصنوعی (همانند ایلان ماسک، استیو ورنیاک و دیگران) به خاطر برخی نگرانی‌ها (در سال ۲۰۲۳) با امضای نامه‌ای سرگشاده خواستار توقف موقت تحقیقات هوش مصنوعی شدند تا زمینه تدوین مقررات و قوانین فراهم شود. در همین سال، سام آلتمن مدیرعامل «OpenAI» در مقابل کنگره آمریکا به نگرانی‌ها از ناحیه هوش مصنوعی اقرار کرد و خواستار نظارت بر هوش مصنوعی شد (<https://www.techtarget.com/searchenterpriseai/definition/AI-governance>).

چند سال پیش نیز بیست‌هزار محقق نامه‌ای را امضا کرده‌اند و به سیاست‌گذاران درباره خطرات سلاح‌های خودمختار - که ممکن است از الگوریتم‌هایی برای حمله به اهداف خاص استفاده کنند - هشدار دادند (وی‌یوان و کنگ، ۲۰۱۸، ص ۳).

افزون بر این، در حال حاضر ابتکارات زیادی در زمینه‌های اخلاقی و اجتماعی و جنبه‌های حقوقی هوش مصنوعی در سراسر دنیا و در سازمان‌های بین‌المللی و مراکز دانشگاهی و تحقیقاتی در حال انجام است. برای نمونه، محققان هوش ماشینی مؤسسه «MIRI» هدف خود را «همسوسازی هوش مصنوعی پیشرفته با منافع انسانی» عنوان کرده‌اند (<https://intelligence.org>).

استنفورد «مؤسسه هوش مصنوعی انسان‌محور» را در سال ۲۰۱۹ تأسیس کرد که هدف اصلی آن پیشبرد تحقیقات، آموزش و سیاست‌گذاری هوش مصنوعی برای بهبود شرایط انسانی است (<https://hai.stanford.edu>). مؤسسه مهم دیگر «مؤسسه آینده زندگی» (Future of Life Institute) است که در سال ۲۰۱۵ تأسیس شده است. مأموریت این مؤسسه هدایت فناوری تحول‌آفرین به سمت سودمندی و دوری از خطرات بسیار شدید است. (<https://futureoflife.org>).

از دیگر مراکز در این حوزه، «رصدخانه سیاست هوش مصنوعی» (OECD AI Policy Observatory) است. این رصدخانه خود را یک مرکز جهانی برای سیاست‌گذاری هوش مصنوعی معرفی کرده است. این مرکز سعی دارد با تبیین برخی اصول به کشورها کمک کند تا خطرات هوش مصنوعی را مدیریت کرده، از منافع آن بهره‌مند شوند (<https://oecd.ai/en/wonk>). یکی از کارهای مهم این رصدخانه آن است که بیش از ۷۰۰ ابتکار سیاست ملی هوش مصنوعی را از ۶۰ کشور و منطقه ثبت کرده است (جوناس و دیگران، ۲۰۲۳، ص ۷-۸).

اما علت این همه نگرانی و توجه به جنبه‌های حقوقی، حکمرانی و اخلاقی از ناحیه فناوری هوش مصنوعی چیست؟ برخی از ابعاد چرایی نگرانی‌های اخلاقی را در ادامه بیشتر تبیین می‌کنیم.

حجم وسیعی از داده‌ها ممکن است از ناحیه هوش مصنوعی دستکاری شود. این امر می‌تواند در یادگیری ماشینی و نحوه ارتباط افراد با دیگران تأثیر بگذارد؛ یعنی هوش مصنوعی می‌تواند در چگونگی فکر و استدلال فرد به‌عنوان شخص انسانی تأثیر بگذارد. این دست کارها همچنین می‌تواند پیامدهایی را برای سازمان‌دهی بازار کار داشته باشد (بادینگتون، ۲۰۱۷، ص ۲-۳). در همه موارد مذکور، ممکن است هوش مصنوعی در عملکردها و جهت‌گیری‌ها تأثیر بگذارد. بنابراین نگران‌های مهم حقوقی و اخلاقی از ناحیه هوش مصنوعی برای انسانیت وجود دارد.

البته نگرانی‌ها و خطرات بسیاری از ناحیه هوش مصنوعی و در حوزه‌های گوناگون قابل طرح است. برخی نگرانی می‌تواند جامعه‌شناختی و آینده‌پژوهانه باشد؛ همچون مسئله فنی و طغیان هوش مصنوعی و ناتوانی برای مهار ابرهوش. همچنین برخی مسائل مهم حکمرانی، سیاسی و امنیتی باید در نظر گرفته شود؛ همانند سوء استفاده قدرت‌های استخباراتی برخی دولت‌ها و سازمان‌های بین‌المللی از هوش مصنوعی برای مهار، نفوذ، حمله و یا استعمار سایر کشورها. برخی چالش‌های جدی در حوزه‌های معرفت‌شناختی، روان‌شناختی، انسان‌شناختی و دین‌شناختی را هم می‌توان بررسی کرد؛ همچون تأثیراتی که هوش مصنوعی بر تغییر شناخت و رفتار انسان‌ها می‌گذارد و به تبع آن، اراده، تصمیم‌گیری، رفتار و فرهنگ افراد دستخوش تغییرات خواهد شد.

این مقاله بیشتر بر نگرانی‌های اخلاقی از ناحیه خود هوش مصنوعی تمرکز کرده است. از این‌رو به‌منظور پایش و هدایت این فناوری باید یک سلسله بایدها و نبایدهای اخلاقی مد نظر قرار گیرد.

به سبب این نگرانی‌ها و خطرات، اخیراً «یونسکو» به فکر افتاده است چارچوب‌های نظارت بین‌المللی و ملی را فراهم آورد تا این فناوری در نهایت، به نفع بشریت تمام شود. همان‌گونه که مقدمه سند اخلاقی «یونسکو» به این

مسئله تصریح می‌کند، هرچند فناوری هوش مصنوعی می‌تواند خدمت بزرگی برای بشریت به‌شمار آید و ملت‌ها از آن بهره‌مند شوند، اما نگرانی‌های اساسی و اخلاقی نیز وجود دارد. در کنار تهدیدهای زیستی و اقتصادی، تهدیدهای فرهنگی نیز ممکن است از این ناحیه متوجه جوامع بشری گردد؛ یعنی هوش مصنوعی ممکن است بر کرامت انسانی، حقوق بشر، آزادی، سیاست و فرهنگ تأثیرگذار باشد. در بندهای ۱۳، ۱۴، ۱۵ و ۱۶ آمده است: کرامت انسانی، حقوق افراد، و آزادی‌های اساسی انسانی باید از ناحیه هوش مصنوعی حفظ شود و آسیب نبیند.

همچنین در بخش C بند ۲ توصیه‌نامه به برخی مسائل جدید اخلاقی که از ناحیه هوش مصنوعی شکل می‌گیرد، اشاره شده است. برای نمونه، هوش مصنوعی در تصمیم‌گیری، اشتغال و کار، تعامل اجتماعی، مراقب‌های بهداشتی، آموزش، رسانه، دسترسی به اطلاعات، شکاف دیجیتالی، داده‌های شخصی و حمایت از مصرف‌کننده، محیط‌زیست، مردم‌سالاری، حاکمیت و قانون، امنیت، آزادی بیان، حریم خصوصی، نفی تبعیض و عدالت تأثیر می‌گذارد. افزون بر این، ظرفیت هوش مصنوعی برای انجام برخی کارها ممکن است انحصار انجام برخی رفتارها در موجود زنده و حتی در انسان را به پایان برساند. از این‌رو در این سند آمده است: امکان دارد در درازمدت سامانه‌های هوشمند خودآگاهی انسان، تعامل اجتماعی، فرهنگی و محیطی، استقلال، عامل بودن و حتی احساس خاص انسان‌ها از تجربه و عامل بودن را به چالش بکشد.

همچنین در بند ۵ و ۱۲ آمده است: این توصیه‌نامه با هدف ایجاد زمینه‌ای برای کارکرد سامانه‌های هوش مصنوعی به نفع بشریت، افراد جوامع و محیط‌زیست برای جلوگیری از آسیب، ارائه می‌شود و در پی تشویق به استفاده صلح‌آمیز از این فناوری است. همچنین در بندهای ۲۵ و ۲۷ برای جلوگیری از آسیب زدن هوش مصنوعی به انسان‌ها تدابیری اندیشه شده است.

بنابراین هوش مصنوعی می‌تواند چالش‌های اخلاقی بسیاری داشته باشد. افزون بر این، حتی در این سند ذکر شده است که به‌طور کلی هوش مصنوعی می‌تواند در علوم انسانی تأثیر بگذارد.

در بند ۳ این سند آمده است: «یونسکو» به مباحث آموزش، علم، فرهنگ و ارتباطات توجه ویژه دارد و در تبیین معنای «علم» در بخش B همین بند آورده است: مراد از «علم» معنای گسترده‌ای است که از علوم طبیعی و پزشکی گرفته تا علوم اجتماعی و انسانی را شامل می‌شود. علت تأثیر آن را هم این‌گونه بیان می‌کند: از طریق فناوری هوش مصنوعی، ظرفیت‌ها و رویکردهای تحقیقاتی جدیدی ایجاد می‌شود و به تبع آن، بر تصورات ما از درک و توضیح عملی تأثیر می‌گذارد. از این‌رو هوش مصنوعی می‌تواند مبنای جدیدی برای تصمیم‌گیری ما ایجاد کند.

به سبب همین تأثیرات، رهبر معظم انقلاب تأکید کرده است فرهنگ و دین اسلامی ایران باید در پیشرفت‌های علمی و فناوری، اصل قرار گیرد، نه اینکه فرع بر آنها باشد. همچنین در همه نوع پیشرفت (از جمله پیشرفت در حوزه فناوری) می‌بایست لوازم فرهنگی آن مد نظر مسئولان قرار گیرد و پیوست فرهنگی داشته باشد (حسینی خامنه‌ای، ۱۳۹۲). افزون بر این، به سبب اهمیت فناوری و تأثیر بر جنبه‌های گوناگون زندگی انسان، اخیراً رهبر حکیم انقلاب در حوزه هوش مصنوعی تأکید کرد:

پیشنهاد می‌کنم یکی از مسائلی که مورد تکیه و توجه و تعمیق واقع می‌شود، مسئله هوش مصنوعی باشد که در اداره آینده دنیا نقش خواهد داشت... باید کاری کنیم که حداقل به ده کشور اول دنیا در این مسئله برسیم (حسینی خامنه‌ای، ۱۴۰۰).

در پی این بیانات، برخی سازمان‌ها و نهادهای همچون «شورای عالی انقلاب فرهنگی» و «معاونت علمی و فناوری ریاست جمهوری» به تکاپو افتادند تا برنامه‌ای در حوزه پیشرفت هوش مصنوعی تدوین کنند. اما به نظر می‌رسد در جامعه ما بیشتر جنبه فنی - مهندسی هوش مصنوعی مد نظر قرار گرفته است.

به بیان دیگر، به نظر می‌رسد به‌طور کلی دو رویکرد بین مهندسان و فلاسفه و اندیشمندان نسبت به مقوله «هوش مصنوعی» وجود دارد:

یک رویکرد نگرش ابزاری صرف به سامانه‌های هوش مصنوعی است. در این نگرش، جنبه‌های فنی - مهندسی و استفاده ابزاری آن بیشتر مد نظر است.

رویکرد دیگر، رویکرد بنیادین، فلسفی و تحلیلی به مسئله هوش مصنوعی است.

در رویکرد اول نگرش خوش‌بینانه به این فناوری وجود دارد و هرچند ممکن است به برخی مضرات و خطرات آن نیز اشاره شود، ولی بیشتر جنبه‌های فنی - مهندسی و استفاده از منافع و کاربردهای این فناوری مد نظر است. اما در رویکرد دوم با اعتراف به اینکه هوش مصنوعی منافع برای جوامع دارد، خطرات و نگرانی‌های مهم اخلاقی نیز از ناحیه فناوری هوش مصنوعی متوجه انسان‌هاست. افزون بر این، در رویکرد بنیادین، فناوری هوش مصنوعی دارای مبانی و مسائل فلسفی، حکمرانی، فرهنگی و تمدنی بوده، همچنین می‌تواند در این حوزه‌ها نیز تأثیر بسزایی بگذارد. از این‌رو در این پژوهش رویکرد دوم، یعنی رویکرد تحلیلی - فلسفی به فناوری هوش مصنوعی مد نظر قرار گرفته است تا بتوان به صورت اساسی این موضوع را بررسی کرد.

خلاصه اینکه به‌سبب این تأثیرات مهم از ناحیه هوش مصنوعی بر مبانی علوم انسانی - به‌طور عام - و بر ارزش‌های اخلاقی - به‌طور خاص - «یونسکو» درصدد برآمده است به‌عنوان بازوی فرهنگی سازمان ملل، ارزش‌های اخلاقی را در تعامل با هوش مصنوعی وضع کند.

بنابراین مسئله اصلی و اساسی پژوهش حاضر این است که ارزش‌های اخلاقی سند مذکور مبتنی بر کدام مکتب اخلاقی است؟ آیا این مبانی با اصول ارزش‌های اخلاقی مکتب اسلام و فیلسوفان مسلمان، همچون آیت‌الله مصباح یزدی سازگاری دارد؟

۲. مفهوم‌شناسی

پیش از ورود به بررسی سند و مسائل مزبور، شایسته است به اجمال اصطلاح «اخلاق هوش مصنوعی» (Ethics of Artificial Intelligence) و رویکردهای مختلف در این زمینه تبیین شود. هرچند اصطلاح «اخلاق هوش مصنوعی» به‌تازگی مطرح شده، اما واژه «اخلاق» در ارتباط با انسان از دیرباز در میان

اندیشمندان مسلمان و غربی محل بحث قرار گرفته است. از این رو شایسته است برای تبیین مفهوم «اخلاق هوش مصنوعی»، نخست تبیین مختصری درباره مفهوم «اخلاق» ارائه دهیم:

۱-۲. اخلاق

«اخلاق» جمع «خُلُق» است. علمای اخلاق اسلامی تعریف‌های متنوعی برای «خُلُق» ارائه داده‌اند. در تعریفی، «خُلُق» ملکه‌ای است که نفس به مقتضای آن، به‌سادگی فعلی را بدون احتیاج به فکر کردن و تأمل نمودن انجام می‌دهد. فرق «حال» با «ملکه» آن است که «حال» کیفیت نفسانی است که به‌سرعت از بین می‌رود، ولی «ملکه» کیفیتی نفسانی است که به‌سختی از بین می‌رود (طوسی، ۱۴۱۳ق، ص ۶۳).

همان‌گونه که صاحبان صنعت و حرفه به‌سهولت کار خود را انجام می‌دهند، صاحبان ملکات نفسانی نیکو و زشت نیز به‌سهولت کار خوب یا بد انجام می‌دهند. افزون بر این، در تعریف «علم اخلاق» در کتاب *فلسفه اخلاق* آیت‌الله مصباح یزدی آمده است:

می‌توان آن را به داناتی تعریف کرد که از انواع صفات خوب و بد و چگونگی اکتساب صفات خوب و زدودن صفات بد بحث می‌کند. موضوع علم اخلاق عبارت است از: صفات خوب و بد از آن جهت که مرتبط با افعال اختیاری انسان بوده، برای او قابل اکتساب یا اجتناب‌اند (مصباح یزدی، ۱۳۹۱، ص ۲۵).

از این رو بنا بر نگرش آیت‌الله مصباح یزدی، اخلاق تنها ناظر به حوزه صفات درونی انسان نیست، بلکه رفتار اختیاری را هم که منجر به کسب این صفات می‌شود، دربر می‌گیرد. اما در سوی دیگر، به باور برخی اندیشمندان علم اخلاق مغرب‌زمین (نظیر ژکس) ارزش‌های اخلاقی تنها ناظر به رفتار افراد است. از این رو آنها علم اخلاق را بیان‌کننده ارزش و لزوم رفتارها معرفی می‌کنند (ژکس، ۱۳۶۲، ص ۹).

همچنین از منظر مکاتب غیرواقع‌گرا و قراردادگرا - که در ادامه دیدگاه آنها تبیین خواهد شد - تمرکز اصلی بر روی رفتار انسان است. در این صورت، از منظر آنها هدف اصلی اخلاقی تصحیح خود رفتار است و ورای آن هدف دیگری را دنبال نمی‌کنند (مصباح، ۱۳۹۰، ص ۲۶).

اما در توصیه‌نامه «یونسکو» تعریف دقیق و روشنی از «اخلاق» ارائه نشده است. البته می‌توان به‌طور کلی بیان کرد: هدف این توصیه از تأکید بر ارزش‌های اخلاقی، اصلاح صفات درونی انسانی نیست. دلیل این ادعا به مبانی هستی‌شناختی، انسان‌شناختی و ارزش‌شناختی و روح سند مذکور برمی‌گردد - که در ادامه، تبیین خواهد شد.

افزون بر این، برخی مفاد این توصیه، همچون بند اول تصریح می‌کند که هدف اخلاقی این سند اصلاح برخورد و رفتار مسئولانه است و در کل سند، به صفات درونی انسانی اشاره‌ای نشده است. بنابراین مشخص می‌شود که تمرکز اصلی این سند بر تصحیح رفتار انسان‌هاست.

هرچند این دغدغه در جای خود ارزشمند است، اما اگر تنها تصحیح رفتار بدون توجه به باطن انسان معیار باشد، با مبانی اخلاق اسلامی فاصله دارد.

اما در اندیشه علامه مصباح یزدی، هم رفتار انسانی از اهمیت خاصی برخوردار است و هم صفات نفسانی. وی افزون بر رویکرد برخی اندیشمندان مسلمان - که تنها به صفات اخلاقی توجه دارند - بر موضوع رفتار نیز تأکید دارد. از منظر ایشان، رفتار اخلاقی است که موضوع قضیه اخلاقی قرار می‌گیرد (مصباح یزدی، ۱۳۸۴، ص ۳۲۵-۳۲۶). بنابراین در مباحث اخلاق، هم صفات اکتسابی باید مد نظر قرار گیرد و هم رفتارهای اختیاری؛ زیرا هم صفات درونی (مثل سخاوت، بخل، تواضع و کبر) متصف به خوبی و بدی و متعلق باید و نباید قرار می‌گیرد و هم رفتارهای بیرونی انسان نیز قابل تحسین و تقبیح است و متعلق باید و نباید قرار می‌گیرد. همچنین بیشتر صفات و ملکات اخلاقی از تکرار اعمال و رفتار افراد به وجود می‌آید.

در نتیجه، تفاوت اساسی این دو رویکرد این است که اگر صفات درونی و ملکات انسانی مورد توجه قرار گیرد، شما می‌توانید حالات درونی (مثل نیت، اخلاص، ریا و از این قبیل) را ارزش‌گذاری کنید. ولی اگر هدف و معیار ارزش اخلاقی تنها تصحیح رفتار انسان باشد، دیگر نمی‌توانید حالات درونی انسان را که ظهور بیرونی ندارد، ارزیابی کنید. بنا بر این نگرش - برای نمونه اگر کسی با نیت خالص و الهی برای خدمت به مردم مدرسه‌ای بسازد و دیگری با نیت ناخالص و یا شیطانی و یا تنها برای معروف‌شدن مدرسه‌ای بسازد، هر دو رفتار به یک میزان ارزشمند است؛ زیرا هر دوی آنها رفتار صحیح را - که همان مدرسه‌سازی است - به‌درستی انجام داده‌اند، اما با انگیزه و نیت متفاوت که در این رویکرد قابل سنجش نیست.

افزون بر این، ناظر به مسئله پژوهش حاضر، پیامد مهم دیگر این دو رویکرد مشخص می‌شود. اگر تنها ملاک رفتار باشد، می‌توان رفتار هوش مصنوعی را ارزش‌گذاری اخلاق کرد و حتی با رفتار و موجود مختاری مانند انسان مقایسه و ارزیابی کرد. فرض کنید یک ربات خدمتکار یا همدم طراحی شده است که از سالمند و یا کودک مراقبت می‌کند. حال اگر روزی رفتاری انجام داد و به سالمند و کودک آسیبی زد، بنا بر رفتارگرایی می‌توان رفتار این ربات را ارزش‌گذاری اخلاقی کرد.

ولی اگر افزون بر رفتار، در انجام افعال، صفات درونی هم مطرح باشد، مسئله به این سادگی نیست؛ زیرا باید بررسی کرد که آیا تصمیم‌گیری ربات بر انجام این فعل ناشی از آگاهی، اراده، و اختیار است؟ یا اینکه ربات‌ها و هوش مصنوعی حالات درونی (مثل آگاهی و اراده) همانند انسان ندارند تا آن فعل را «اختیاری / اخلاقی» بدانیم، بلکه این نوع تصمیم‌گیری سامانه هوشمند ممکن است ناشی از مشکلات سخت‌افزاری و یا نرم‌افزاری تعبیه‌شده در آن باشد؟

۲-۲. اخلاق هوش مصنوعی

آیا منظور «از اخلاق هوش مصنوعی» این است که خود ماشین‌های هوشمند رفتار اخلاقی داشته باشند؟ یا اینکه منظور این است که انسان‌ها در استفاده از هوش مصنوعی، ارزش‌های اخلاقی را مراعات کنند؟ به‌طور کلی، می‌توان گفت: بنا بر مبانی هستی‌شناختی و انسان‌شناختی، رویکردهای متفاوتی در این مسئله می‌تواند وجود داشته باشد.

۳. رویکردهای مختلف در اخلاق هوش مصنوعی

۳-۱. رویکرد اول

در یک رویکرد، خود هوش مصنوعی و ربات‌ها مخاطب باید و نباید اخلاقی قرار می‌گیرد. «قوانین سه‌گانه رباتیک» که نخستین بار در سال ۱۹۴۲ توسط ایزاک آسیموف برای جلوگیری از آسیب‌رسانی ربات‌ها به انسان ارائه شد، از این قسم است (ر.ک. راسل و نورویگ، ۲۰۲۰).

۳-۲. رویکرد دوم

در رویکرد دیگر، اخلاق هوش مصنوعی بخشی از اخلاق ماشین است؛ یعنی هدف از اخلاق ماشین این است که می‌بایست ماشینی ایجاد شود که دست‌کم از یک اصل اخلاقی مطلوب و یا مجموعه‌ای از اصول اخلاقی پیروی کند. (اندروسن و اندروسن، ۲۰۰۷، ص ۱۵).

مقصود از «پیروی هوش مصنوعی از اصل اخلاقی» دارای ابهام است. آیا منظور این است که ربات به اختیار خودش یک اصل اخلاقی را دنبال کند؟ اگر این نگرش مراد باشد، همان رویکرد اول خواهد بود. ولی اگر منظور این است که سامانه‌های هوشمند طوری برنامه‌ریزی شود که رفتار اخلاقی متناسب با رفتار انسان از آنها سر بزنند، این را می‌توان رویکرد دیگری غیر از رویکرد اول به حساب آورد.

به عبارت واضح‌تر، در رویکرد دوم منظور از «اخلاق هوش مصنوعی» این است که چگونه می‌توان رمزهای (کدهای) اخلاقی را در سامانه‌های هوشمند تدوین کرد تا از خطرات احتمالی هوش مصنوعی علیه انسان‌ها جلوگیری شود. (بادینگتون، ۲۰۱۷، ص ۱) البته خطرات از ناحیه هوش مصنوعی منحصر به انسان‌ها نیست، بلکه می‌تواند به حیوانات و محیط‌زیست نیز آسیب برساند.

به هر حال، در این رویکرد دیگر خود ماشین‌های هوشمند متصف به باید و نباید و یا تصمیم‌گیری اخلاقی نمی‌شوند. بلکه آنها به‌نوعی از جبر اخلاقی برخوردارند؛ زیرا آنها به‌مثابه یک ماشین حساب می‌شوند؛ البته ماشین هوشمندی که برنامه‌های اخلاقی داده‌شده به آنها را به‌خوبی پیاده‌سازی می‌کند.

۳-۳. رویکرد سوم

اما در رویکرد متفاوت‌تر و عام‌تر از دو رویکرد پیشین، اخلاق هوش مصنوعی ناظر به چگونگی استفاده انسان‌ها از این ابزار و فناوری نوین است؛ یعنی انسان‌ها به طور مستقل مخاطب اصلی باید‌ها و نباید‌های اخلاقی هستند. در این رویکرد انسان‌ها مسئول رفتار خویش‌اند و همه نقش‌آفرینان حوزه هوش مصنوعی - اعم از دولت‌ها، سازمان‌ها، مهندسان و برنامه‌نویسان، کاربران و استفاده‌کنندگان - را دربر می‌گیرد. از این‌رو همه این نقش‌آفرینان موظف‌اند در جهت ارزش‌های اخلاقی این فناوری را توسعه دهند یا از آن استفاده کنند.

تا اینجا به اختصار سه رویکرد درباره اخلاق هوش مصنوعی تبیین شد، اما در ادامه، بیشتر به تحلیل و بررسی این سه رویکرد می‌پردازیم. در تحلیل این سه رویکرد لازم است این مسائل بررسی شود: آیا خود ربات‌ها می‌توانند به باید و نباید

اخلاقی متصف شوند؟ (تحلیل رویکرد اول) آیا بنا بر رویکرد دوم از منظر اخلاق اسلامی، می‌توان برنامه‌ریزی را برای سامانه‌های هوشمند ارائه کرد؟ (تحلیل رویکرد دوم) سند اخلاق هوش مصنوعی ناظر به کدام‌یک از این سه رویکرد نوشته شده است؟ و از منظر مبانی اخلاق اسلامی بنا بر رویکرد سوم چه نظریه‌ای را می‌توان ارائه داد؟ (تحلیل رویکرد سوم)

پاسخ به این مسائل و اتخاذ رویکرد مختار از بین رویکردهای سه‌گانه، در گرو آن است که بررسی شود محققان و اندیشمندان چه نگرشی به هویت هوش مصنوعی دارند.

۴. هوش مصنوعی قوی و ضعیف

بنا بر رویکرد اول خود سامانه‌های هوشمند به باید و نباید اخلاقی متصف می‌شدند. تحلیل و نقد این رویکرد و اتخاذ رویکرد مناسب در این‌باره نیازمند آن است که دو اصطلاح رایج در حوزه هوش مصنوعی توضیح داده شود. به نظر می‌رسد این نوع رویکردهای متفاوت در حوزه اخلاق هوش مصنوعی، متأثر از دو مکتب فکری مختلف به مقوله هوش مصنوعی باشد:

یک مکتب «رویکرد هوش مصنوعی ضعیف» (Weak AI) است. این مکتب فکری با مؤسسه فناوری ماساچوست مرتبط است. در این رویکرد تنها معیار این است که برنامه به کاررفته در هوش مصنوعی (AI) به‌درستی و رضایتمندانه انجام شود. اما «رویکرد هوش مصنوعی قوی» (Strong AI) مکتب فکری دیگری است که با رویکرد دانشگاه کارنگی ملون به هوش مصنوعی ظهور پیدا کرده است. در این مکتب رفتار هوش مصنوعی با رفتار انسان سنجیده می‌شود. این در حالی است که در رویکرد ضعیف با عملکرد سامانه سنجیده می‌شد. در رویکرد قوی به دنبال شبیه‌سازی سامانه ادراکی انسان در هوش مصنوعی هستند. افزون بر این، طرفداران هوش مصنوعی قوی بر این باورند که به سبب دستیابی به اکتشافات، الگوریتم‌ها و دانش برنامه‌های نوین، هوش مصنوعی می‌تواند دارای حس، هوشیاری و هوش باشد (لوسی و کوپس، ۲۰۱۶، ص ۱۲).

به‌طور کلی در رویکرد قوی، رفتار ماشین بر اساس شباهت با عملکرد انسان سنجیده می‌شود. اما در رویکرد ضعیف مهندس تنها انتظار رفتار موفقیت‌آمیز از سامانه هوشمند طراحی شده را دارد.

با این وصف، در توصیه‌نامه اخلاقی هوش مصنوعی «یونسکو» به‌صراحت به دو رویکرد قوی و ضعیف در زمینه هوش مصنوعی اشاره نشده، اگرچه در برخی بندهای سند از آسیب‌ها و خطرات پیشرفت ماشین‌های هوشمند در آینده اعلام نگرانی شده است. از این‌رو به نظر می‌رسد برخی نگرانی‌ها ناظر به رویکرد هوش مصنوعی قوی در این فناوری وجود دارد. برای نمونه در انتهای بخش C بند ۲ توصیه‌نامه - که به برخی نگرانی‌های اخلاقی از ناحیه هوش مصنوعی اشاره شده - آمده است که ظرفیت هوش مصنوعی برای انجام برخی کارها، انحصار انجام برخی رفتارها در موجود زنده و حتی در انسان را به پایان می‌رساند. از این‌رو در این سند آمده: ممکن است در درازمدت سامانه‌های هوشمند خودآگاهی انسان، تعامل اجتماعی و فرهنگی و محیطی، استقلال، عامل بودن و حتی احساس خاص انسان‌ها از تجربه و عامل بودن را به چالش بکشد.

بنابراین در اسناد بین‌المللی اگرچه بیشتر بر هوش مصنوعی ضعیف تأکید شده، ولی در عبارت مزبور آمده است: عاملیت انسان و یا خودآگاهی انسان و یا احساس انسان بودن ممکن است در آینده دچار چالش گردد. زمانی این امور به چالش کشیده می‌شود که هوش مصنوعی نیز دارای این ویژگی‌ها شود. به بیان دیگر، با پیشرفت هوش مصنوعی ممکن است این فناوری هوشمند به حوزه‌های اختصاصی انسانی ورود پیدا کند و همانند انسان ویژگی‌هایی نظیر احساسات و خودآگاهی داشته باشد. بدین‌سان در این سند با بیان این دغدغه‌ها و نگرانی‌ها به‌نوعی به امکان تحقق هوش مصنوعی قوی در آینده اشاره شده است.

به‌هرحال، آیا بنا بر مبانی علم‌النفس اسلامی، ربات‌ها می‌توانند به باید و نباید اخلاقی متصف شوند و یا همانند انسان تصمیم‌های اخلاقی بگیرند؟ انجام رفتار اخلاقی متوقف بر آگاهی و درک معنای «خوب» و «بد» و مبتنی بر اختیار و اراده است. اراده نیز مستلزم داشتن ذهن و یا نفس است. ازاین‌رو بنا بر مبانی علم‌النفس اسلامی، هوش مصنوعی (قوی) زمانی می‌تواند رفتاری اخلاقی از خود نشان دهد که دارای ذهن و یا نفس خودآگاه باشد. هرچند بررسی این مسئله و امکان تحقق هوش مصنوعی قوی نیازمند یک تحقیق مستقل است و پرداختن به آن در حوصله این مقاله نیست، ولی در ادامه به اختصار این مسئله را بررسی می‌کنیم:

هرچند کارکردهای پیچیده‌ای از سامانه هوشمند بروز می‌کند و در برخی موارد (همچون محاسبه، پردازش اطلاعات، حافظه و یادگیری) ممکن است از انسان‌های معمولی قوی‌تر باشد، ولی نمی‌توان آن را به‌مثابه یک «ساختار واحد زیستی» و به‌عنوان یک «هویت واحد، زنده و خود آگاه» به‌حساب آورد، بلکه هوش مصنوعی به لحاظ وجودی، یک ماشین و ربات مرکب از اجزاست که یک سلسله کارکردهای هماهنگ انجام می‌دهد. ازاین‌رو بنا بر مبانی علم‌النفس صدرایی شیئی همچون هوش مصنوعی به لحاظ وجودی، هویت واحد ندارد. پس نمی‌تواند ذهن و نفس داشته باشد. افزون بر این، به نظر می‌رسد در توصیه‌نامه اخلاقی «یونسکو» با تعریفی که از «هوش مصنوعی» ارائه شده، رویکردی ضعیف اتخاذ شده است. همان‌گونه که در بند ۲، به‌ویژه بخش A تأکید می‌کند، نگاه این سند به سامانه‌های هوش مصنوعی به‌مثابه فناوری‌های دارای پردازش اطلاعات، به شیوه‌ای شبیه رفتار هوشمند است.

همچنین در بند ۲۶ اشاره می‌کند: هوش مصنوعی نباید تصمیم‌گیر نهایی در مسائل مهمی همچون مرگ و زندگی باشد، بلکه اختیار آن باید به دست انسان‌ها باشد. این بدان معناست که هوش مصنوعی در انجام افعال خود اختیار تام ندارد. این مطلب ناظر به رویکرد هوش مصنوعی ضعیف است؛ زیرا اگر منظور هوش مصنوعی قوی بود خودش می‌توانست به‌تنهایی تشخیص دهد و تصمیم بگیرد و مسئولیت رفتار خویش را نیز بپذیرد. افزون بر این، مراد از «اخلاق هوش مصنوعی» در این توصیه‌نامه متمرکز بر ارتباط با رفتار انسان‌ها و بازیگران هوش مصنوعی است، نه اینکه خود سامانه‌های هوشمند را مخاطب باید و نباید اخلاقی قرار دهد.

۵. تحلیل سه رویکرد در اخلاق هوش مصنوعی

بنابراین رویکرد اول به امکان اخلاقمندی خود ربات‌ها همانند انسان، بنا بر مبانی علم‌النفس اسلامی متنی است. به بیان دیگر، خود هوش مصنوعی نمی‌تواند به باید‌ها و نبایدهای اخلاقی متصف شود.

اما رویکرد دوم امکان اخلاق‌مندی ربات و هوش مصنوعی به لحاظ مهندسی و برنامه‌نویسی، نه اینکه امکان‌پذیر باشد، بلکه ضروری است. همان‌گونه که در مقدمه و برخی بندهای توصیه‌نامه «یونسکو» آمده است، باید هوش مصنوعی را به سمت ارزش‌های اخلاقی هدایت کرد. در این زمینه بعضی دانشگاه‌ها، مراکز تحقیقاتی و مؤسسات در تلاش‌اند تا برای کاهش خطرات هوش مصنوعی، رمزهای اخلاقی را در آنها تعبیه کنند (بادینگتون، ۲۰۱۷، ص ۳). البته آنها مبتنی بر مبانی و ارزش‌های اخلاقی خود سعی دارند تا با رمزنویسی، هوش مصنوعی را اخلاق‌مند کنند. اما جامعه اسلامی مبانی و ارزش‌های اخلاقی خاص خود را دارد. بنابراین با توجه به رویکرد دوم در امکان اخلاق‌مندی ماشین‌های هوشمند در جامعه اسلامی می‌توان این نظر را مطرح کرد: «می‌توان با به کار بردن الگوریتم‌ها و با ارائه داده‌ها و برنامه‌هایی مبتنی بر ارزش‌های اخلاقی اسلامی، به گونه‌ای رفتار ماشین را جهت‌دهی کرد. همچنین از سامانه‌های هوشمند در ترویج ارزش‌های اخلاقی اسلامی و مقابله با مطالب ضد اخلاقی استفاده نمود».

لازم به ذکر است به سبب آنکه حوزه اخلاق هوش مصنوعی یک حوزه میان‌رشته‌ای است، این نظر زمانی محقق خواهد شد که همکاری و هم‌افزایی بین مهندسان و برنامه‌نویسان هوش مصنوعی با اندیشمندان حوزه علوم انسانی و علوم دینی و فلاسفه به صورت جدی و اساسی شکل بگیرد.

تا اینجا در تبیین مفهوم «اخلاق هوش مصنوعی» سه رویکرد ارائه شد و مشخص گردید بنا بر مبانی اسلامی، رویکرد اول در امکان اخلاق‌مندی خود سامانه‌های هوشمند و ربات‌ها متغی است. اما اخلاق‌مندی هوش مصنوعی بنا بر رویکرد دوم، نه تنها امکان‌پذیر است، بلکه ضروری هم هست؛ چنان که در سطح دنیا این مسئله مورد اهتمام قرار گرفته است. از این رو در جامعه اسلامی نیز باید با همکاری مهندسان و اندیشمندان حوزه اخلاق اسلامی این امر محقق شود. اما رویکرد سوم در اخلاق هوش مصنوعی به بایدها و نبایدهای استفاده انسان از این فناوری برمی‌گردد. این رویکرد در بخش پیش رو و در ضمن تحلیل منشأ ارزش اخلاقی سند «یونسکو» بررسی و تبیین خواهد شد. در انتها بنا بر مبانی فلسفه اخلاق آیت‌الله مصباح یزدی، نظریه‌ای درباره اخلاق هوش مصنوعی در ارتباط با رفتار انسان ارائه خواهد گردید.

۶. تحلیل منشأ ارزش اخلاقی توصیه‌نامه و سازگاری آن با نظریه اخلاقی اسلام

منشأ ارزش و لزوم اخلاقی به کار برده شده در این توصیه‌نامه با کدام یک از مکاتب اخلاقی سازگاری دارد؟ به نظر می‌رسد منشأ ارزش و هنجارهایی که در این توصیه‌نامه به آنها اشاره می‌شود از نوع قراردادگرایی باشد. این مطلب با تأمل در برخی مفاد قابل اثبات است. در مقدمه سند، حقوق بشر و آزادی را به تأیید و پذیرش مردم مبتنی کرده است. همچنین در بند ۶ بیان نموده است: «هدف این توصیه‌نامه رسیدن به یک ابزار هنجاری پذیرفته شده در سطح جهانی است». در نتیجه، این مفاهیم را ملاک هنجارهای اخلاقی معرفی می‌کند و بر اساس آنها «بایدها و نبایدهای اخلاقی» را در حوزه هوش مصنوعی معین می‌سازد. عبارت‌های مذکور ما را به این سمت سوق می‌دهد که بپذیریم مبانی به کار رفته در این متن، برخاسته از مکاتب غیرواقع‌گرایی و از نوع قراردادگرایی است.

اما اشکال قراردادگرایی چیست؟ (نگارنده در تحقیق دیگری به تفصیل این مطلب را تبیین کرده است. ر.ک. محسنی و دیگران، ۱۴۰۲، ص ۱۹۴-۲۰۴). اشکال اصلی آن است که مکتب «قراردادگرایی» ذیل مکاتب غیرواقع‌گرا می‌گنجد (ر.ک. مصباح، ۱۳۹۰؛ مصباح یزدی، ۱۳۸۴). از این‌رو نسبیت‌گرایی و کثرت‌گرایی از لوازم مهم قراردادگرایی است (ر.ک. مصباح، ۱۳۹۰؛ مصباح یزدی، ۱۳۸۴).

با این وجود، مفاد بندهای ۱۳۱ تا ۱۳۴ که در حوزه نظارت و ارزیابی تأثیرات اخلاقی فناوری‌های هوش مصنوعی است و سعی دارد روشی را برای این ارزیابی ارائه دهد، دچار مشکل اساسی می‌شود؛ زیرا بنا بر کثرت‌گرایی، دیگر نمی‌توان روش واحدی را برای بررسی تأثیرات اخلاقی در کشورهای عضو ارائه داد.

ممکن است در پاسخ این اشکال که گفته شود: جوامع بر یک اصل کلی و مشترک اجماع کرده‌اند و آن اصل کلی نظیر نفی ضرر هوش مصنوعی می‌تواند ملاک ارزیابی باشد.

در نقد این پاسخ می‌توان گفت: این اصل ثابت است یا در جوامع مختلف متفاوت است؟ اگر این اصل ثابت و مطلق باشد به نوعی ناقض دیدگاه کثرت‌گرایی خواهد بود و اگر در جوامع مختلف فرق کند باز نمی‌تواند ملاک واحدی برای ارزیابی همه جوامع باشد.

۷. بررسی موضوع و هدف علم اخلاق در مکاتب اخلاقی

اما مسئله اصلی که در این کاوش به آن می‌پردازیم، تبیین موضوع و هدف علم اخلاق است. در فلسفه اخلاق دیدگاه‌های متفاوتی درباره موضوع و هدف اخلاقی وجود دارد. با تبیین برخی دیدگاه‌ها درباره موضوع و هدف علم اخلاق، رویکرد این توصیه در این زمینه نیز روشن می‌شود.

اگر بخواهیم درباره چیزی بیندیشم فرق می‌کند که با چه سؤال‌هایی آغاز کنیم. پرسش‌های اساسی ارسطو در اخلاق نیکوماخوس درباره شخصیت است. ارسطو با این پرسش شروع می‌کند: «خیر بشری چیست؟» و پاسخ او این است: خیر بشری فعالیتی از روح آدمی است که مطابق با فضیلت باشد. بنابراین برای فهم اخلاق باید فهمید که چه چیزهایی شخص را بافضیلت می‌کند (ریچلز، ۱۳۸۹، ص ۲۳۹).

علاوه بر رفتار، ما درباره اشخاص هم داوری می‌کنیم. اخلاق مبتنی بر فضیلت عمدتاً با اشخاص سروکار دارد؛ اما اخلاق مبتنی بر فضیلت به راهنمایی کردار هم می‌پردازد. حکم تعیین‌کننده در اخلاق مبتنی بر کردار این است که «کاری را درست انجام بده». حکم تعیین‌کننده در اخلاق فضیلت‌محور چیزی نیست جز اینکه «شخص خوبی باشد» (هولمز، ۱۳۸۹، ص ۷۸).

همان‌گونه که گذشت، ارزش‌های اخلاقی در این سند مبتنی بر نظریه «قراردادگرایی» است. نظریه «قراردادگرایی» ذیل دیدگاه «وظیفه‌گرایی اخلاقی» (Ethics of Duty) می‌گنجد. در این دیدگاه معیار ارزیابی اخلاقی عمل به تکلیف و وظیفه است (محیطی و احمدی، ۱۴۰۰، ص ۱۶۲). نظریه «اخلاق وظیفه» دیدگاهی است

که از سوی برخی از اندیشمندان علم اخلاق مغرب‌زمین مطرح شده است. ارزش‌های اخلاقی را صرفاً ناظر به رفتارها می‌دانند و علم اخلاق را بیان‌کننده ارزش و لزوم رفتارها معرفی می‌کنند (ژکس، ۱۳۶۲، ص ۹). هدف اصلی اخلاقی را نیز تصحیح خود رفتار دانسته، و رای آن هدف دیگری را نشان نمی‌دهند (مصباح، ۱۳۹۰، ص ۲۶).

اما در دیدگاه «اخلاق فضیلت» (Ethics of Virtue) صفات درونی نفسانی متصف به ارزش‌های اخلاقی می‌شود و علم اخلاق عهده‌دار بیان ارزش و لزوم این صفات است. هدف علم اخلاق در این رویکرد، آراستن نفس به کمالات نفسانی پسندیده و زودن صفات ناپسند از آن است. رفتارها و اعمال انسان که وسیله‌ای برای تحقق این امر است نیز در این رویکرد ارزش‌گذاری می‌شود. به عبارت دیگر، هدف در این رویکرد دستیابی به کمال نفس است. برخی علمای مسلمان بر اساس انسان‌شناسی خاص خود، همین دیدگاه را اختیار کرده‌اند. از نظر ایشان حقیقت وجود انسان نفس مجرد اوست و کمال و نقص واقعی انسان نیز ناظر به همین نفس یا روح است. کمال و نقصی که با افعال اختیار انسان در این دنیا به دست می‌آید در روح او باقی می‌ماند و انسان پس از مرگ با همان‌ها محشور می‌شود (نراقی، ۱۲۰۹ق، ج ۱، ص ۳۶-۳۹).

در رویکرد «فضیلت‌محور» ارزش‌گذاری مربوط به اشخاص است، نه رفتار. اما این تأکید در فلسفه اخلاق نوین جای خود را به رویکردی به شدت قانون‌پرستانه، به رفتار داده است. از این‌رو قواعد و اصول بر صدر نشسته‌اند (هولمز، ۱۳۸۹، ص ۷۰). درواقع وجه تمایز اصلی در این است که اخلاق فضیلت‌محور ناظر به ارزش‌های اخلاقی بوده و دیدگاه وظیفه‌گرا ناظر به الزام اخلاقی است. از این‌رو مسئله اصلی در اخلاق فضیلت‌محور این است که «چگونه انسانی باشیم؟» و در نظریه وظیفه‌محور دغدغه اصلی این است که «چه فعلی را باید انجام دهیم؟» (خزاعی، ۱۳۸۹، ص ۱۸). بنابراین یک شخص ممکن است از فضایل اخلاقی بی‌بهره باشد، ولی به قصد انجام وظیفه، یک عمل اخلاقی را انجام دهد.

به همین سبب است که برخی ارزش‌گذاری‌هایی که انجام می‌دهیم مربوط به اشخاص است، نه کردار. ما حکم می‌کنیم که مردم خوب یا بد، وظیفه‌شناس یا بی‌مسئولیت، درخور ستایش یا درخور نکوهش هستند. این رویکرد به جای آنکه تأکیدش بر کردار باشد، در پی راهنمایی انسان به جانب صفاتی است که بزرگ‌منشی را به بار می‌آورند. این صفات را «فضایل» می‌نامند و این جهت‌گیری معرفت اخلاق مبتنی بر فضیلت است (هولمز، ۱۳۸۹، ص ۷۰).

به باور اخلاق فضیلتی‌ها، به سبب فضیلت راست‌گویی است که نباید دروغ گفت؛ یعنی این نظریه بر راست‌گو بودن تأکید دارد و آن دو (سودگرایی و وظیفه‌گرایی) بر راست گفتن. وجه تمایز اصلی آنها این است که اخلاق فضیلت ناظر به ارزش‌های اخلاقی است و دو نظریه «سودگرایی» و «وظیفه‌گرایی» ناظر به الزام اخلاقی. از این‌رو مسئله اصلی در اخلاق فضیلت این است که «چگونه انسانی باشیم؟» و در دو نظریه دیگر اینکه «چه فعلی را باید انجام دهیم؟» تمایز بین فضیلت‌محوری و عمل‌محوری موجب می‌شود تا بتوانیم بین انجام یک عمل به ظاهر فضیلت‌مندانه و فضیلت‌مندانه عمل نمودن تمایز قائل شویم (خزاعی، ۱۳۸۹، ص ۱۸).

از این رو نقد مبنایی به متن «یونسکو» آن است که مفاد توصیه‌نامه به علت آنکه از قسم مکاتب قراردادگرایی غیرواقع‌گراست، وظیفه‌محور و عمل‌محور است. پس همان‌گونه که در بند ۱۰ سند «یونسکو» آمده، ارزش‌های ذکرشده در این سند الهام‌بخش رفتار مطلوب است؛ یعنی ارزش‌های اخلاقی آن تنها متوجه اصلاح کنش و رفتار ظاهری انسان است. در نتیجه، در این سند اخلاقی اصلاً به مباحثی همچون کمالات نفسانی، رابطه انسان با خدا، تأثیر واقعی ارزش‌های اخلاقی در روح و نفس انسان و مواردی از این دست، اشاره نشده است.

اما نظریه اخلاقی مکتب اسلام با رویکرد آیت‌الله مصباح یزدی بسی متفاوت از دیدگاه فضیلت‌گرایانه است. این دیدگاه نکات مثبت دیدگاه «فضیلت‌گرایی اسلامی» را در خود دارد و افزون بر آن، نوعی دیدگاه غایت‌انگارانه و کمال‌جویانه است. در این اندیشه، رفتار اختیاری نیز موضوع علم اخلاق است (مصباح یزدی، ۱۳۸۴، ص ۳۲۵)؛ زیرا برخی صفات نفسانی با تکرار رفتار اختیاری به دست می‌آید (مصباح یزدی، ۱۳۷۳، ص ۹). رفتار اختیاری انسان وسیله‌ای برای رسیدن به اهداف مطلوب است. از این رو رفتار به خاطر وسیله و مقدمه بودن، دارای ارزش است (مصباح یزدی، ۱۳۸۸، ج ۱، ص ۲۶۲). بنابراین در رویکرد آیت‌الله مصباح یزدی در این اندیشه، هم رفتار اختیاری مد نظر است و هم صفات نفسانی. در این نظریه مفاهیم ارزش‌های اخلاقی مبتنی بر واقعیت است. از این رو این مکتب ذیل نظریه «واقع‌گرا» قرار می‌گیرد.

اهدافی که انسان‌ها برای تحقق بخشیدن به آنها می‌کوشند، یا تأمین نیازمندی‌های طبیعی و پاسخ‌گویی به نیازها و غرایز است و یا تأمین منافع و مصالح اجتماعی و جلوگیری از فساد و هرج‌ومرج. ولی هدف اصلی انسان رسیدن سعادت ابدی و کمال معنوی و روحی است (مصباح یزدی، ۱۳۸۸، ج ۱، ص ۲۶۲).

اما به نظر می‌رسد هدف اخلاقی این سند در رویکرد خوش‌بینانه، تأمین مصالح اجتماعی است. ولی در نظریه اخلاقی مکتب اسلام و آیت‌الله مصباح یزدی، هدف رسیدن به کمال نهایی است. مطلوبیت هدف به لحاظ حیثیت تکوینی امر عینی و مستقل از میل و رغبت و تشخیص افراد است. هرگاه فعل را در ارتباط با هدف مطلوب از جهت مطلوبیت آن در نظر بگیریم مفهوم «ارزش» از آن انتزاع می‌شود، و هرگاه رابطه وجودی با نتیجه مترتب بر آن لحاظ شود مفهوم «وجوب» یا «شایستگی» و یا «بایستگی» از آن گرفته می‌شود که در اندیشه فلسفی از آن به «ضرورت بالقیاس» تعبیر می‌گردد. به بیان دیگر بین فعل اختیاری و نتایج مترتب بر آن رابطه سبب و مسببی برقرار است (مصباح یزدی، ۱۳۸۸، ج ۱، ص ۲۶۳-۲۶۴).

آیت‌الله مصباح یزدی در تفکیک بین موضوعات حقوقی و اخلاقی به نکته‌ای مهم اشاره کرده است که می‌تواند به بحث حاضر و تمایز اخلاق «وظیفه‌گرا» و «کمال‌جو» نیز مربوط شود:

در چنین مواردی اگر رعایت این قاعده فقط به انگیزه ترس از مجازات دولت باشد، ارزش اخلاقی ندارد، هرچند کاری است موافق با موازین حقوقی؛ و اگر به انگیزه هدف‌عالی‌تر که همان هدف اخلاقی است انجام گیرد، کاری اخلاقی هم خواهد بود (مصباح یزدی، ۱۳۸۸، ج ۱، ص ۲۶۷).

همچنین دیدگاه آیت‌الله مصباح یزدی به‌نوعی دیدگاه غایت‌انگارانه و کمال‌جویانه به معنای خاص خود است. هدف از فعل اختیاری قرب وجودی به کمال مطلق و خداوند متعال است. در این نظریه مبتکرانه میان فعل اختیاری انسان و نتیجه آن رابطه ضرورت بالقیاس برقرار است. از این‌رو ارزش اخلاقی فعل اختیاری تابع تأثیر آن فعل در رسیدن انسان به کمال حقیقی است. پس هر کاری به‌اندازه‌ای که در رسیدن به کمال حقیقی انسان مؤثر باشد، ارزنده خواهد بود، و اگر تأثیر منفی داشته باشد ارزش آن نیز منفی خواهد بود، و اگر هیچ ارزش مثبت یا منفی نداشته باشد خنثا خواهد بود (مصباح یزدی، ۱۳۸۴، ص ۳۲۶-۳۲۷).

بنا بر نظریه اخلاقی مکتب اسلام درباره ارزش‌های اخلاقی هوش مصنوعی، می‌توان این نظریه را ارائه داد: «اگر چرخه سامانه‌های هوش مصنوعی، از اختراع و کشف گرفته تا کاربرد و استفاده آن، در راستای کمال انسانی و تقرب به خداوند متعال باشد، این کار ارزش اخلاقی دارد و به‌عکس، اگر فناوری هوش مصنوعی موجب دوری انسان از کمال حقیقی خود شود، ضدارزش به‌حساب می‌آید. در غیر این صورت، آن فناوری خنثا بوده و دارای ارزش و یا ضدارزش اخلاقی نیست». بنابراین رویکرد سوم اخلاق هوش مصنوعی در ارتباط با انسان از منظر اخلاق اسلامی مشخص شد و معیار و بایدها و نبایدهای پیشرفت و استفاده از این فناوری در جهت اخلاق اسلامی و تقرب به خداوند ارزش‌گذاری می‌شود.

نتیجه‌گیری

تحقیق حاضر به لحاظ مبانی، سند اخلاق هوش مصنوعی «یونسکو» را بررسی کرد. با این کاوش مشخص شد سند مذکور مبتنی بر قراردادی‌گرایی است. در این مکتب هدف عمل به وظیفه و اصلاح رفتار است، نه کسب صفات کمالی. اما نظریه اخلاقی اسلام جزو مکاتب واقع‌گرایانه بوده و در اندیشه برخی علمای فضیلت‌محور و از منظر آیت‌الله مصباح یزدی غایت‌گرا و کمال‌جوست. هدف نهایی در اندیشه این فیلسوف اخلاق، افزون بر تصحیح رفتار، ساختن شخص انسان و تقرب به خداوند متعال است.

از این‌رو پیشنهاد می‌شود در تنظیم قراردادهای بین‌المللی، بررسی اسناد، اعلام نظرها و توافق‌ها دقت لازم شود و سازگاری آن با مبانی اندیشه اسلامی لحاظ گردد. همچنان‌که رهبر معظم انقلاب می‌فرماید، در بحث یادگیری فناوری، شاگردی کردن و علم‌آموزی اشکالی ندارد؛ اما نباید به توصیه و برنامه‌های آنها خوش‌بین بود و عمل کرد (حسینی خامنه‌ای، ۱۳۹۷).

افزون بر این، پیشنهاد می‌شود در نگاشت اسناد و تدوین قوانین و مقررات ملی در حوزه هوش مصنوعی، متناسب با ارزش‌ها و چارچوب‌های اخلاقی و آموزه‌های اسلامی زمینه‌سازی شود.

منابع

- نگمارک، ماکس (۱۴۰۱ق). زندگی ۳/۱۰/انسان بودن در عصر هوش مصنوعی، ترجمه میثم لطفی، چ هشتم، تهران: فرهنگ نشر نو - آسیم.
- جمعی از نویسندگان (۱۴۰۲). نقد و تکمیل پیش نویس نهایی توصیه نامه درباره اخلاق هوش مصنوعی، دفتر فقه معاصر، قم: مرکز مدیریت حوزه های علمیه.
- حسینی خامنه ای، سیدعلی (۱۳۹۲). سخنرانی در دیدار با اعضای شورای عالی انقلاب فرهنگی. ۱۳۹۲/۹/۱۹.
- حسینی خامنه ای، سیدعلی (۱۳۹۷). سخنرانی در دیدار با جمعی از مسئولان نظام، سفرای کشورهای اسلامی و قشرهای مختلف مردم، ۱۳۹۷/۱۱/۳.
- حسینی خامنه ای، سیدعلی (۱۴۰۰). سخنرانی در دیدار با جمعی از نخبگان و استعدادهای برتر علمی کشور. ۱۴۰۰/۸/۲۶.
- خداقلیزاده، بهزاد و دیگران (۱۳۹۹). نقد و بررسی اولین پیش نویس توصیه نامه اخلاق هوش مصنوعی (یونسکو). تهران: پژوهشگاه فضای مجازی.
- خزاعی، زهرا (۱۳۸۹). اخلاق فضیلت. تهران: حکمت.
- ریچلز، جمیز (۱۳۸۹). فلسفه اخلاق. ترجمه آرش اخگری. تهران: حکمت.
- ژکس (۱۳۶۲). فلسفه اخلاق (حکمت عملی). ترجمه ابوالقاسم پورحسینی. چ دوم، تهران: امیرکبیر.
- طوسی، نصیرالدین (۱۴۱۳ق). اخلاق ناصری. تهران: علمیه اسلامی.
- محسنی، حمید و دیگران (۱۴۰۲). بررسی توصیه نامه اخلاقی هوش مصنوعی یونسکو و تبیین نقاط ضعف و قوت آن. جمعی از نویسندگان، نقد و تکمیل پیش نویس نهایی توصیه نامه درباره اخلاق هوش مصنوعی، دفتر فقه معاصر، قم: مرکز مدیریت حوزه های علمیه.
- محسنی، حمید و فولادی قلعه، کاظم (۱۴۰۱). سازگاری و ناسازگاری توصیه نامه اخلاق هوش مصنوعی یونسکو با نظریه اخلاقی اسلامی، اولین کنفرانس فضای سایبر، قم: دانشکده مهندسی دانشکدگان فارابی دانشگاه تهران.
- محیطی، حسن و احمدی، حسین (۱۴۰۰). فلسفه اخلاق، فرااخلاق و اخلاق هنجاری. قم: مؤسسه آموزشی و پژوهشی امام خمینی.
- مصباح یزدی، محمدتقی (۱۳۷۳). دروس فلسفه اخلاق. تهران: اطلاعات.
- مصباح یزدی، محمدتقی (۱۳۸۴). نقد و بررسی مکاتب اخلاقی، تحقیق و نگارش احمدحسین شریفی، چ سوم، قم: مؤسسه آموزشی و پژوهشی امام خمینی.
- مصباح یزدی، محمدتقی (۱۳۸۸). آموزش فلسفه، چ هشتم، تهران: سازمان تبلیغات اسلامی.
- مصباح یزدی، محمدتقی (۱۳۹۱). فلسفه اخلاق، تحقیق و نگارش احمدحسین شریفی، قم: مؤسسه آموزشی و پژوهشی امام خمینی.
- مصباح، مجتبی (۱۳۹۰). بنیاد اخلاق (روشی نو در آموزش فلسفه اخلاق)، چ هفتم، قم: مؤسسه آموزشی و پژوهشی امام خمینی.
- نراقی، مهدی (۱۲۰۹ق). جامع السعادات. تحقیق محمد کلانتر، بیروت: مؤسسه الاعلمی للمطبوعات.
- هولمز، رابرت ال (۱۳۸۹). مبانی فلسفه اخلاق. ترجمه مسعود علیا، تهران: ققنوس.

Anderson, M. & Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*. 28(4), 15-26.

Boddington P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. New York: Springer International Publishing.

- Dafoe, A. (2018). *AI Governance: A research agenda*, Oxford: Oxford university press.
- Jonas, T. & et al, (2023). *The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research*. Retrieved from <https://arxiv.org/ftp/arxiv/papers>.
- Lucci, S. & Kopec, D. (2016). *Artificial intelligence in the 21st century: A living introduction, mercury learning and information*. ISBN-13: 978-1942270003, ISBN-10: 1942270003.
- Russell, S. & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. 4th Ed. USA: Prentice Hall.
- UNESCO (2021). Recommendation on the ethics of artificial intelligence. Retrieved from URL: <https://en.unesco.org/artificial-intelligence/ethics>.
- Weiyu W. & Keng S. (2018). *Artificial Intelligence: A Study on Governance, Policies, and Regulations*, MWAIS 2018 Proceedings.40. Retrieved from <http://aisel.aisnet.org/mwais2018/40>.
<https://hai.stanford.edu>.
<https://oecd.ai/en/wonk>.
<https://unesdoc.unesco.org>.
<https://www.techtarget.com>.
<https://futureoflife.org/resource/ai-policy>.

