Hekmat va Falsafeh
Vol. 20, No.78, Summer 2024, Pp 45-73
wph.atu.ac.ir
DIO: 10.22054/wph.2024.75680.2183

Original Research
Received:10/09/2023 Accepted: 21/05/ 2024.
eISSN: 2476-6038
ISSN: 1735-3238

Convergence of Privacy and Transparency, Limitations of Artificial Intelligence Design

Mohammad Ali Ashouri Kisomi *

PhD of Philosophy of Art, Allameh Tabataba'i University, Tehran, Iran

Abstract

This research aims to criticize the approach that considers the solution to the ethical challenges of artificial intelligence (AI) to be only in design and technical improvements. Some researchers consider the ethical challenges in AI to be convergent, which emerge with the advent of AI systems and are resolved with technical progress and improvements. In the discussions of the ethics of AI, issues such as privacy and transparency have been the focus of most studies. In the present research, using the analytical-critical method, the convergence between transparency and privacy in machine learning systems was investigated, and the approach that limits the resolution of ethical challenges to AI design was criticized. The research findings indicate that there is no convergence between the ethical challenges of AI. Additionally, by raising three challenges of quantification, technical limitations, and meta-ethical issues, the criticism of approaches that rely solely on design and technical improvements was addressed. The results also indicate that it is not possible to respond to the ethical challenges of AI only by relying on design, and there is a need to use other methods such as legislation or progress in other sciences in addition to attention to design.

Keywords: artificial intelligence ethics, privacy, transparency, opacity, ethical design.

* m_ashori@atu.ac.ir

How to Cite: Ashouri Kisomi, Mohammad Ali. (2024). Convergence of Privacy and Transparency, Limitations of Artificial Intelligence Design, *Hekmat va Falsafeh*, 20 (78), 45-73.
DIO: 10.22054/wph.2024.75680.2183

1. Introduction

One of the challenges in the ethics of artificial intelligence (AI) is the actual implementation of laws and human rights. Issues such as identifying legal responsibility, providing proof, etc. still lie ahead of us. Although significant efforts have been made in this direction, such as establishing the European Union's GDPR, many problems remain. Among the existing views on the ethics of AI, some believe that the solution to these problems should be sought in technical matters. For example, proponents of 'Internet Libertarianism' or 'Cyber Libertarianism' can be cited. These views, to some extent, are based on the assumption that technical solutions can solve the social problems created by AI systems. According to this approach, the ethical issues and challenges of AI are somewhat convergent. In other words, the emergence of AI brings about ethical issues; these problems are intensified during the development of AI systems (convergence in increasing ethical challenges); and finally, AI and its technical reforms solve these problems (convergence in resolving ethical challenges). On the contrary, the main goal of this research is to examine and investigate this view, focusing on the two ethical difficulties of privacy and transparency in AI systems.

To achieve the main goal of the research, the relationship and convergence of privacy and transparency in machine learning systems will be examined. Following the rejection of convergence, reasons will also be provided against approaches that consider the ethical challenges of AI in design and technical reforms. In line with reaching the intended goals, the research is divided into four sections. In the initial section, what we mean by AI and machine learning will be introduced. After clarifying the main concepts, the second section examines ethical approaches to privacy in the ethics of AI. The third section deals with transparency/opacity, its types, and its relationship with privacy. In the fourth section, after identifying the challenges ahead in the previous sections, the relationship between transparency and privacy is examined, and criticisms are presented.

Research Question(s)

1. In the context of machine learning systems, to what extent are the ethical considerations of privacy protection and transparency convergent?

2. Literature Review

Numerous studies have been conducted on the ethics of artificial intelligence and, more precisely, on machine learning systems, and a very rich and extensive body of research literature is forming. In this regard, privacy and opacity/transparency are topics that have repeatedly garnered the attention of philosophers. Some believe that without

transparency, privacy cannot be protected; as we will lack a rationale for the machine's judgments, which are considered essential for protecting privacy. Consequently, with increased opacity/decreased transparency, privacy will be more at risk.

3. Methodology

The current study employs an analytical-critical methodology to explore the nexus between transparency and privacy within machine learning systems. Furthermore, this study challenges the limited perspective that ethical issues in AI can be addressed only through the way AI systems are ethically designed.

4. Conclusion

Considering the objectives mentioned, the findings of this research are divided into two parts. In the first part, the results indicate that in the discussions on the ethics of artificial intelligence, there is no necessity for convergence between transparency/opacity and privacy. In other words, increased transparency does not guarantee the protection of privacy. It is noteworthy that if these two issues are considered convergent, the result could be the endangerment of privacy. Sometimes transparency and the protection of privacy in an AI system may improve together, but this is not a general rule and should not be interpreted as one strengthening the other. It was also determined that there are various methods to protect privacy in AI systems, and we do not always need transparency. These results implicitly support the idea that we should consider the ethical discussions of artificial intelligence as separate and non-convergent issues. This lack of convergence shows that we are dealing with significant issues, not just one ethical issue and its subsets. In the second part, the research results demonstrate that although, in many cases, technical solutions and design adjustments are effective, they face limitations in some situations and will not be a solution for all circumstances. Just as philosophical and ethical approaches that do not consider the various structures of artificial intelligence can lead to general analyses that often have no relation to the structures of AI; if we consider technical solutions as general solutions without regard to different conditions, we will again make a mistake in another way. In such situations, it is necessary to use other methods, such as legislation, advancements in other domains of science, along with ethical design, to overcome some of the existing limitations.

مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۶/۱۹ تاریخ پذیرش: ۱۴۰۳/۰۳/۰۱

ISSN :۱۷۳۵-۳۲۳۸ eISSN: ۲۴۷۶-۶۰۳۸

حکمت و فلسفه

دورهٔ ۲۰، شمارهٔ ۷۸، تابستان ۱۴۰۳، صص ۴۵-۷۳

wph.atu.ac.ir

DIO: 10.22054/wph.2024.75680.2183

همگرایی حریم خصوصی و شفافیت، محدودیت‌های طراحی هوش مصنوعی

محمدعلی عاشوری کیسمی* دکتری فلسفه هنر دانشگاه علامه طباطبائی، تهران، ایران

چکیده

هدف از این پژوهش نقد به رویکردی است که راهکار برطرف شدن چالش‌های اخلاقیِ هوشِ مصنوعیِ را محدود به طراحی و اصلاحات فنی می‌داند. برخی پژوهش‌گران چالش‌های اخلاقی در هوش مصنوعی را همگرا تلقی می‌کنند و معتقدند این چالش‌ها همانطور که با ظهور سیستم هوش مصنوعی پدید آمدند، با پیشرفت و اصلاحات فنی آن مرتفع خواهند شد. در مباحثِ اخلاقِ هوش مصنوعی، موضوعاتی همچون حفاظت از حریم خصوصی و شفافیت در بیشتر پژوهش‌ها موردتوجه قرار گرفته است. در پژوهش حاضر با استفاده از روشِ تحلیلی-انتقادی، همگرایی میان شفافیت و حریم خصوصی در سیستم‌های یادگیری ماشین بررسی شده و رویکردی که رفع چالش‌های اخلاقی را به طراحی هوش مصنوعی محدود می‌داند به نقد گذاشته شده است. نتایج این بررسی نشان می‌دهد که همگرایی میان چالش‌های اخلاقی هوش مصنوعی وجود ندارد. همچنین با مطرح کردن سه چالش کمی‌سازی، محدودیت‌های تکنیکی و مباحث فرااخلاقی به نقد رویکردهایی که تنها بر طراحی و اصلاحات فنی متکی هستند پرداخته شد. نتایج این بخش نشان می‌دهد که نمی‌توان تنها با تکیه‌بر طراحی، به چالش‌های اخلاقی در هوش مصنوعی پاسخ داد و نیاز به استفاده از سایر روش‌ها مانند قانون‌گذاری و یا پیشرفت در سایر علوم در کنار توجه به طراحی وجود دارد.

واژه‌های کلیدی: اخلاق هوش مصنوعی، حریم خصوصی، شفافیت، کدری، یادگیری ماشین، طراحی اخلاقی.

m_ashori@atu.ac.ir

۱. مقدمه

با گسترش استفاده از هوش مصنوعی[۱] مباحث اخلاقی آن موردتوجه فیلسوفان قرار گرفت. دسته‌ای از این مباحث اخلاقی بر اساس تصوراتی است که امروزه ازلحاظ فنی امکان‌پذیر نیست (Hagendroff & Wezel, 2020)؛ اما بسیاری از آن‌ها، ممکن است در آینده به حقیقت بپیوندند (Muller & Bostrom, 2016). زمانی که از هوش مصنوعی سخن به میان می‌آید عموماً پژوهش‌ها به ده مبحث اصلی اخلاقی تمرکز می‌کنند[۲]: حریم خصوصی و نظارت[۳]، دستکاری در رفتار[۴]، کدری[۵] و شفافیت[۶]، سوگیری در تصمیمات سیستم[۷]، اثر متقابل انسان-ربات[۸]، اتوماسیون و اشتغال[۹]، سیستم‌های خودمختار[۱۰]، اخلاق ماشین[۱۱]، عامل اخلاقی مصنوعی[۱۲] و تکینگی[۱۳] (Muller, 2021). این مباحث گاهی با یکدیگر تداخل پیدا کرده و نمی‌توان همواره مرز دقیقی میان آن‌ها مشخص کرد. این مباحث معمولاً در سیستم‌های مختلف هوش مصنوعی خود را نمایان می‌سازند. به عبارتی دقیق‌تر، مباحث اخلاقی فوق ممکن است در برخی سیستم‌های هوش مصنوعی دارای اهمیت بود و در برخی سیستم‌ها اصولاً چندان قابل‌تأمل نباشند[۱۴]. همچنین، باید توجه داشت شکل بروز و

---

۱. Artificial Intelligence

۲. البته بسیاری از تاثیرات خارجی هوش مصنوعی کمتر مورد توجه قرار می‌گیرد و در پژوهش‌های جریان اصلی معمولاً به موضوعاتی پرداخته می‌شود که راهکارهای فنی برای آن‌ها وجود دارد (Hagendroff, 2022).

۳. Privacy and surveillance

۴. Manipulation of behavior

۵. Opacity

۶. Transparency

۷. Bias in decision systems

۸. Human-Robot interaction

۹. Automation and Employment

۱۰. Autonomous Systems

۱۱. Machine Ethics

۱۲. Artificial Moral Agents

۱۳. Singularity

۱۴. به‌عنوان مثال، اثرات متقابل انسان-ربات در یک سیستم نرم‌افزاری هوش مصنوعی دارای اهمیت نخواهد بود. یا در یک سیستم هوش مصنوعی ضعیف، مبحث تکینگی محلی از بحث پیدا نخواهد کرد.

توجه به هرکدام از این مباحث می‌تواند با توجه به سیستم‌های مشخص با یکدیگر متفاوت باشد[۱]. در میان تکنیک‌های مختلف هوش مصنوعی، یکی از مهم‌ترین آن‌ها از نظر فیلسوفان و پژوهشگرانِ اخلاق، «یادگیریِ ماشین[۲]» است، اما زمانی که از یادگیری ماشین سخن به میان می‌آید نیز تنها با یک شیوه یادگیری روبرو نیستیم (Bhatnagar, 2019). پژوهش‌های فراوانی در خصوص اخلاق هوش مصنوعی در یادگیری ماشین صورت گرفته است و ادبیات پژوهشی بسیار غنی و گسترده‌ای در حال شکل‌گیری است. در این میان حریم خصوصی و کدری/شفافیت از جمله مباحثی است که به کرات موردتوجه فیلسوفان قرار گرفته است. به‌صورت خلاصه کدری/شفافیت به مشخص بودن/نبودن روندی که سیستم به یک نتیجه مشخص می‌رسد اشاره دارد (Muller, 2021a). برخی معتقدند بدون شفافیت نمی‌توان از حریم خصوصی حفاظت کرد؛ چراکه توجیهی برای قضاوت‌های ماشین که لازمه حفاظت از حریم خصوصی است در دست نخواهیم داشت. بر این اساس، با افزایش کدری/کاهش شفافیت، حریم خصوصی بیشتر در خطر خواهد بود (ر. ک Muller, 2021; Franzoni, 2023 Barhamgi & Bertino, 2022;). زمانی که با این قبیل چالش‌های اخلاقی روبرو می‌شویم، یک راه‌کار استفاده از قوانین و مقررات است.

یکی از چالش‌هایی که در حوزه هوش مصنوعی و تکنولوژی‌های دیجیتال وجود دارد، اجرای واقعی قوانین است. مشکلاتی همچون شناسایی مسئولیت حقوقی، اثبات و محکمه‌ای صالح که بتوان در صورت قصور به آن مراجعه کرد و همچنین اجرایی کردن تصمیمات این محکمه همچنان پیش روی ما است. اگرچه در این مسیر تلاش‌های فراوانی همچون وضع مقررات عمومی حفاظت از داده اتحادیه اروپا[۳] و پیش از آن قانون حفاظت

---

۱. به‌عنوان مثال زمانی که از سوگیری در یک سیستم هوش مصنوعی نسبتاً ساده که تنها اطلاعات افراد محدودی را نگهداری کرده و آن را طبقه‌بندی می‌کند صحبت می‌کنیم با زمانی که از سوگیری در یک سیستم پیچیده بر پایه شبکه‌های عصبی که می‌تواند اطلاعات انبوهی را از میان داده‌های فراوان به‌دست آورد سخن می‌گوییم، دامنه مبحث سوگیری بسیار متفاوت خواهد بود.

۲. Machine Learning

۳. The General Data Protection Regulation (GDPR)

از داده اتحادیه اروپا[1] صورت گرفته است اما هنوز مشکلات بسیاری پابرجا است[2]. در این شرایط بسیاری از شرکت‌ها محصولات دیجیتال را به دور از نگرانی‌های اخلاقی به آزمایش گذاشته و یا منتشر می‌کنند (ر.ک Muller, 2021a).

در میان دیدگاه‌های موجود در خصوص اخلاق هوش مصنوعی، برخی بر این عقیده هستند که راه‌کار برطرف شدن این مشکلات را باید در امور فنی جستجو کرد. به‌عنوان نمونه این دیدگاه‌ها می‌توان به طرفداران «لیبرتریانیسم اینترنت[3]» یا «لیبرتریانیسم سایبر[4]» اشاره کرد. این قبیل دیدگاه‌ها، بر اساس این فرض استوار شده‌اند که راه‌کارهای فنی می‌توانند مشکلات اجتماعی که سیستم‌های هوش مصنوعی پدید آورده‌اند را برطرف کنند[5] (Mozorov, 2013; White, 2014). در این راستا پژوهش های مختلفی نیز صورت گرفته که به‌صورت ضمنی می‌توان آن‌ها را در راستای تقویت این رویکرد در نظر گرفت (ر.ک Lee et al., 2022; Liu et al., 2022; Sengan et al., 2022). بر اساس این رویکردها، مشکلات و چالش‌های اخلاقی هوش مصنوعی با یکدیگر نوعی همگرایی دارند. به‌عبارتی‌دیگر، با پیدایش هوش مصنوعی مشکلات اخلاقی به وجود می‌آیند، در طول توسعه هوش مصنوعی این مشکلات تقویت می شوند (همگرایی در به افزایش چالش‌های اخلاقی) و دست‌آخر هوش مصنوعی و اصلاحات تکنیکی آن این مشکلات را برطرف می‌کنند (همگرایی در رفع چالش‌های اخلاقی). در نقطه مقابل، هدف اصلی این پژوهش بررسی و نقد به این دیدگاه با تمرکز بر دو چالش اخلاقی حریم خصوصی و شفافیت در سیستم‌های

---

۱. 95/46/EC (Data Protection Directive)

۲. گرچه این مبحث از منظر حقوقی دارای ابعاد مختلفی است، اما در پژوهش حاضر قصد ما نه تمرکز بر جنبه حقوقی و سیاست‌گذاری بلکه تمرکز بر جنبه اخلاقی است. لذا از ورود به مباحث حقوقی اجتناب شده است و به دلیل اهمیت، این موضوعات را شایسته بحث‌های بیشتر در پژوهش‌های جداگانه می‌دانیم.

۳. Internet Libertarianism: این اصطلاح را می‌توان تحت عنوان اختیارگرایی اینترنت هم ترجمه کرد. اما از آنجایی که اصطلاح اختیارگرایی ممکن است به معانی دیگری نیز تعبیر شود، از واژه لیبرتریانیسم در متن استفاده شده است.

۴. Cyber Libertarianism

۵. قابل توجه است که لیبرتریانیسم اینترنت یا لیبرتریانیسم سایبر اهدافی همچون آزادی و یا مقابله با پدرسالاری (Paternalism) را در فضای سایبر نیز دنبال می‌کنند (White, 2014). در پژوهش حاضر نیز تلاش بر این نیست که این دیدگاه‌ها به صورت همه‌جانبه مورد بررسی قرار گرفته، رد یا تایید شوند بلکه ما به بخش از این دیدگاه نقد خود را وارد خواهیم کرد که بر طراحی فنی متمرکز است.

هوش مصنوعی است که از تکنیک‌های یادگیری ماشین استفاده می‌کنند.

به‌منظور دستیابی به هدف اصلی پژوهش، ارتباط و همگرایی حریم خصوصی و شفافیت موردبررسی قرار خواهد گرفت. پیرو رد همگرایی، دلایلی نیز در رد رویکردهایی که چالش‌های اخلاقی هوش مصنوعی را در طراحی و اصلاحات فنی می‌داند ارائه خواهد شد. در این مسیر، از روش تحلیلی-انتقادی و توجه به ساختارهای هوش مصنوعی در یادگیری ماشین استفاده خواهیم کرد تا از ارائه تصویر ذهنی و غیرواقعی از این سیستم‌ها پرهیز شود. شایان‌ذکر است، اهمیت پژوهش حاضر در تمرکز بر دو مبحث مشخص در اخلاق هوش مصنوعی برای رد رویکرد یاد شده و بیان برخی از ایرادات به آن است که تاکنون به شکلی منسجم صورت نگرفته است. در راستای رسیدن به اهداف موردنظر، پژوهش به چهار بخش تقسیم شده است. در بخش ابتدایی، آن‌چه از هوش مصنوعی و یادگیری ماشین مراد می‌کنیم معرفی خواهد شد. پس از روشن شدن مفاهیم اصلی، در بخش دوم رویکردهای اخلاقی به حریم خصوصی در اخلاق هوش مصنوعی موردبررسی می‌گیرد. در بخش سوم به بررسی مفهوم شفافیت/کدری، انواع و ارتباط آن‌ها با حریم خصوصی پرداخته می‌شود. در بخش چهارم و پس از مشخص شدن چالش‌های پیش‌رو در بخش‌های پیشین، ارتباط میان شفافیت و حریم خصوصی بررسی شده و انتقادات ارائه می‌شود.

۲. هوش مصنوعی

آغاز آن‌چه امروز به‌عنوان حوزه هوش مصنوعی شناخته می‌شود به کنفرانس دارپا[۱] در کالج دارتموث[۲] باز می‌گردد. البته سخن گفتن از هوش مصنوعی تاریخچه‌ای طولانی‌تر دارد[۳] (Brindsjord & Govindrajulu, 2016). در پژوهش حاضر مقصود ما از هوش

---

۱. DARPA

۲. Dartmouth

۳. برخی پیشینه روش‌شناختی هوش مصنوعی را به فیلسوفانی چون لایب نیتس، هابز و آکویناس باز می‌گردانند (Falsinski, 2016). البته به شکلی جدیدتر اندیشمندانی همچون تیورینگ به هوش مصنوعی توجه شایانی داشتند. تیورینگ آزمایشی مطرح کرد که به‌عنوان آزمایش تیورینگ (Turing Test) شناخته می‌شود. پرسش او این بود که آیا ماشین می‌تواند از نظر زبانی غیرقابل تشخیص از انسان باشد؟ براساس این آزمایش، اگر یک انسان در یک اتاق و یک کامپیوتر در اتاق دیگری باشد، سپس سوالی مشابه به کامپیوتر و انسان داده شود، یک قاضی براساس پاسخ‌های دریافتی باید تشخیص دهد کدام پاسخ مربوط به انسان است. اگر قاضی در ۵۰ درصد موارد، نتواند پاسخ انسان را از

مصنوعی، نوعی سیستمِ محاسباتی است که برای رسیدن به هدف یا اهداف مشخص رفتارهای «هوشمند» اتخاذ می‌کند. در اینجا «هوشمند»، الزاماً هوشمند به‌مانند انسان هوشمند نیست. یک هوش مصنوعی ممکن است توانایی‌های محدودتر یا گسترده‌تری از یک انسان داشته باشد (Muller, 2021a). لذا در این پژوهش مقصود ما هوش مصنوعی به معنای عام نیست و بر هوش مصنوعی در معنای ضعیف متمرکز خواهیم بود[۱].

۱-۲. یادگیری ماشین

همان‌طور که پیش‌تر اشاره شد «یادگیریِ ماشین» یکی از تکنیک‌های مهم هوش مصنوعی است که موردتوجه فیلسوفان قرار گرفته است. دانش یادگیریِ ماشین به مطالعه و بررسی الگوریتم‌ها[۲] و مدل‌هایی[۳] می‌پردازد که با «یادگیری» از «داده‌های ورودی» برای دستیابی به هدف یا اهداف مشخص طراحی می‌شوند. از منظر عملی، برخلاف سیستم‌هایی که از ابتدا برای رسیدن به یک هدفِ مشخص طراحی می‌شوند و با تکرار نتایج مشابه به‌دست می‌دهند، یادگیری ماشین این امکان را فراهم می‌آورد که معماری سیستم با استفاده از تکرار یک فرآیند، تغییر و تطبیق پیدا کند. فرآیندی که در آن تطبیق و تغییر صورت

پاسخ کامپیوتر تمیز دهد، آن‌گاه ماشین موفق به گذراندن آزمایش تیورینگ است (Turing, 1950). البته دکارت هم این موضوع را مطرح کرده بود. به عقیده دکارت اگر ماشینی وجود داشت که می‌توانست از نظر ظاهری و اعمال شبیه به انسان باشد دو راه برای تشخیص ماشین از انسان وجود دارد. براساس روش اول چنین ماشینی قادر نخواهد بود از گفتار و یا نشانه‌های دیگر مانند انسان به شیوه‌ای استفاده کند که منجربه خیر دیگران شود. طبق روش دوم اگر ماشین بتواند برخی از کارها را مانند انسان انجام دهد، این کار را از روی عقل انجام نمی‌دهد و در نتیجه نمی‌تواند تمامی اعمال را مشابه با انسان انجام دهد (Descartes, 1955: 116).

۱. به‌عنوان مثال راسل و نورویگ (۲۰۲۰) معتقدند هوش مصنوعی حوزه‌ای است در تلاش برای ساخت سیستمی که مانند انسان فکر/عمل کند. از آنجایی که اینگونه تعاریف، هوش مصنوعی در معنای قوی (Strong Artificial Intelligence) یا عام (Artificial General Intelligence) را به ذهن متبادر می‌سازد و در زمان نگارش این پژوهش هنوز به چنین مرحله‌ای نرسیده‌ایم، از این قبیل تعاریف دوری شده است.

۲. Algorithms: از منظر تاریخی، آن‌چه امروز به‌عنوان الگوریتم می‌شناسیم ریشه در آثار خوارزمی دارد. پس از او لایب‌نیتس ایده محاسبات نمادین را که امروز جبر رایانه‌ای می‌خوانیم را پروراند. بعدها جورج بول ایده لایب‌نیتس را گسترش داد و جبر بولی را اختراع کرد. دیویت هیلبرت اصول ریاضیاتی را توسعه داد و پایه‌های رسیدن به علوم کامپیوتر امروزی را فراهم کرد (Thomas, 2015: 30-35).

۳. Models

می‌گیرد را «آموزش[1]» می‌نامند؛ به‌عبارتی در یادگیریِ ماشین، «یادگیری» به ماشین آموزش داده می‌شود[2] (El Naqa & Murphy, 2015). اکنون‌که حوزه مشخص پژوهش حاضر

---

۱. Training

۲. یادگیری ماشین را می‌توان ترکیبی از سه جز دانست: داده‌ها، مدل یا فضای فرضیه (Model – Hypothesis space)، تابع ضرر (Loss function). در یادگیری ماشین از مجموعه‌های داده‌ها استفاده می‌کنیم. یک مجموعه داده‌ها از واحدهای اتمی اطلاعات، یا همان داده تشکیل می‌شود که ممکن است به شیوه‌های مختلفی ذخیره شده باشند. مجموعه داده‌ها ممکن است شامل متن، سیگنال‌های دریافتی یک حسگر، فریم‌های ویدئو یا هر چیز دیگری باشند. هر داده دو خصوصیت دارد: ویژگی (Feature) و برچسب (Label). به‌عنوان مثال مجموعه داده‌های یک تصویر می‌تواند شامل ویژگی‌هایی همچون رنگ، تیرگی و روشنی و غیره را شامل شود. برچسب می‌تواند مشخص کند که داده‌های تصویر مربوط به چیست، مثلا یک فرد، یک گل و غیره است. از روی مجموعه داده‌ها ماشین مدل می‌سازد یا یک مدل از پیش ساخته را بهینه می‌کند و با استفاده از تابع ضرر مدل اصلاح می‌شود. در عمل ابتدا داده‌ها جمع‌آوری می‌شوند. از نظر آماری هر چه حجم داده‌ها بیشتر باشد برای یادگیری بهتر است. باید توجه داشت که نمی‌توان این حجم را بی‌نهایت در نظر داشت؛ چراکه منابع سیستم‌ها دارای محدودیت محاسباتی است. پس از جمع‌آوردی داده‌ها، آماده‌سازی داده «Data preparation» انجام می‌شود؛ به‌عنوان مثال ویژگی‌ها و برچسب‌ها ثبت می‌شود و یا ممکن است نیاز باشد داده‌ها را از جهت وجود سوگیری (Bias) بررسی کرد، و یا داده‌ها به‌صورت اتفاقی مرتب شوند تا ماشین از روی ترتیب ثبت داده‌ها، در یادگیری دچار اشتباه نشود. گاهی در این مرحله داده‌ها به دو بخش تقسیم می‌شوند؛ معمولا بخش اعظم آن‌ها برای یادگیری و بخش اندکی برای ارزیابی از هم جدا می‌شوند. ماشین از روی مجموعه داده‌ها به الگوهایی (Pattern) دست می‌یابد که به وسیله آن‌ها می‌تواند مدل بسازد یا یک مدل از پیش موجود را توسعه دهد. مثلا از روی داده‌های تصاویر گل‌ها، ماشین این الگو را می‌یابد که تصاویر گل دارای برچسب گل و ویژگی رنگ قرمز است و این یک الگو اولیه از تصاویر گل است. از مجموع این الگوها ماشین یک مدل از گل را به‌دست می‌آورد. مدلی که به‌عنوان مثال در یک نمودار خط دو بعدی ساده بتواند داده‌های تصاویر گل و سایر تصاویر را از هم جدا کند. با این‌حال مدل‌ها ممکن است دارای اشتباه و نیازمند اصلاحات باشند؛ در این مرحله تابع ضرر وارد می‌شود. تابع ضرر وظیفه بررسی و اصلاح مدل (فرضیه) را دارد. به‌عبارتی میان مدلی که ماشین از روی داده‌ها ساخته و داده‌های واقعی سنجش صورت می‌گیرد تا موفقیت مدل برای پیش‌بینی‌های درست مشخص شود. در ابتدا ممکن است پیش‌بینی‌های ماشین نادرست بوده و نیاز به اصلاح مدل باشد؛ لذا نیاز به تمرین، یادگیری و اصلاح مدل براساس اشتباهات وجود دارد. به زبان ساده ابتدا داده‌های آموزشی وارد مدل می‌شوند/مدل ساخته می‌شود، مدل پیش‌بینی کرده و پیش‌بینی مورد ارزیابی قرار می‌گیرد. داده‌ها به‌روزرسانی می‌شوند و مجددا این روند آن‌قدر تکرار می‌شود تا یادگیری ماشین به مرحله‌ای برسد که پیش‌بینی‌های درست به‌دست دهد. سپس در مرحله ارزیابی، این روند با داده‌های ارزیابی سنجیده می‌شود. در این مرحله از داده‌های ارزیابی برای مدل استفاده می‌شود و در مرحله آزمون تنها موفقیت مدل برای پیش‌بینی بررسی می‌شود (Jung, 2022: 19-39). قابل توجه است که آموزش محدود به داده‌های آموزشی نیست و یک الگوریتمِ خوبِ یادگیریِ ماشین قادر است با پردازش داده‌های جدید و یادگیری از اشتباهات یک یادگیری مادام‌العمر

در هوش مصنوعی معرفی شد، نیاز است مباحث و چالش‌های اخلاقیِ پژوهشِ حاضر (حریم خصوصی و شفافیت/کدری) صریح‌تر مطرح شوند[۱].

۳- حریم خصوصی و یادگیری ماشین

تعاریف گوناگون و متفاوتی برای «حریم خصوصی» ارائه شده و با گزاره‌هایی گوناگونی روبه‌رو هستیم. این گزاره‌ها عموماً در دو گروه گزاره‌های توصیفی و هنجاری دسته‌بندی می‌شوند. گزاره‌های توصیفی، به توصیفِ موقعیت‌ها و شرایطِ حریمِ خصوصی می‌پردازند؛ و گزاره‌های هنجاری به ارزش‌گذاری موقعیت‌ها، شرایط و محدودیت در استفاده از اطلاعات یا پردازشِ اطلاعات می‌پردازند. گزاره‌های هنجاری عموماً به حقِ اخلاقیِ غیرمطلقِ فرد برای کنترل و یا دسترسی به اطلاعات درباره خود، موقعیت‌هایی که دیگران می‌توانند درباره او اطلاعات کسب کنند و تکنولوژی‌هایی که می‌توانند اقدام به تولید، پردازش و یا انتشار اطلاعات او کنند می‌پردازند.

در رویکرد هنجاری، با سه دیدگاه اصلی درباره حریم خصوصی مواجه هستیم: تحویل‌گرا[۲]، غیرتحویل‌گرا[۳] و خوشه‌ای[۴]. تحویل‌گرایان معتقدند که ادعاهای مربوط به حریم خصوصی و ارزش‌های آن مربوط به ارزش‌ها و رویکردهای دیگر اخلاقی مانند امنیت، خودمختاری، دموکراسی، آزادی و غیره است[۵]. در مقابل غیرتحویل‌گرایان معتقدند که حریم خصوصی، ارزشِ ذاتی داشته و نباید آن را به موضوعات دیگر فروکاست[۶]. دسته

داشته باشد (El Naqa & Murphy, 2015). این روند برای تمامی روش‌های یادگیری ماشین به‌صورت مشابه نیست؛ گاهی روندها پیچیده‌تر و گاهی ساده‌تر است.

۱. در این پژوهش هرجا که از هوش مصنوعی یاد می‌شود تنها به روش‌ها و تکنیک‌های یادگیری ماشین اشاره داریم و نه سایر روش‌ها و تکنیک‌های دیگر هوش مصنوعی.

۲. Reductionist: گاهی به تقلیل‌گرا و فروکاهندگان ترجمه می‌شود.

۳. Non-Reductionist

۴. Cluster

۵. به‌عنوان نمونه‌ای از تحویل‌گرایی در حریم خصوصی می‌توان به پژوهش دیوید متیسون (۲۰۰۸) اشاره کرد که در آن رویکرد تحویل‌گرایی توزیعی معرفی می‌شود.

۶. به‌عنوان نمونه از غیرتحویل‌گرایی می‌توان به پژوهش بیت روسلر (۲۰۰۴) اشاره کرد. او با نقد به دیدگاه‌هایی که تنها به فضای عمومی می‌پردازند و فضای خصوصی را به‌عنوان زیرمجموعه آن در نظر می‌گیرند، معتقد است که در نظام لیبرال-دموکراسی حریم شخصی یک ارزش اصلی است که نمی‌توان آن‌را به شیوه‌ی تحویل‌گرایانه فروکاست.

سوم که به‌عنوان رویکرد خوشه‌ای شناخته می‌شود بر این عقیده هستند که خوشه‌ای از قضایای اخلاقی درباره حریم خصوصی وجود دارد و هیچ مؤلفه اساسی واحدی در خصوص حریم خصوصی وجود ندارد[۱]. در طرف دیگر و در رویکرد توصیفی حریم خصوصی از منظر معرفتی موردتوجه قرار می‌گیرند. بر اساس این رویکرد، نداشتن حریم خصوصی به این معنا است که دیگران به اطلاعات خصوصی فرد دسترسی دارند[۲] (Van den Hoven et al., 2020).

زمانی که از حریم خصوصی در یادگیری ماشین سخن به میان می‌آید معمولاً به محافظت از داده‌ها[۳] و امنیت داده‌ها[۴] توجه می‌شود. از منظر فنی راه‌کارهای گوناگونی برای حفاظت از داده‌ها و امنیت آن‌ها وجود دارد. ازجمله این راه‌کارها می‌توان به طراحی ایمن[۵]، کمینه‌سازی داده‌ها[۶]، کنترل دسترسی[۷] و اطلاع‌رسانی اشاره کرد (Jobin et al., 2019). در گذشته تصور بر این بود که اگر داده‌ها توسط ماشین پردازش شوند و مستقیماً در دسترس افراد قرار نگیرند، می‌توان بیشتر از حریم خصوصی محافظت کرد. اما امروزه می‌دانیم که دسترسی ماشین به داده به‌معنای حفاظت از حریم خصوصی نیست (Muller, 2009) و یک واقعیت فنی در این است که بدون وجود داده‌ها نیز نمی‌توان ماشین را آموزش داد.

---

۱. در خصوص رویکرد خوشه‌ای پژوهش‌های گوناگونی صورت گرفته است. به‌عنوان مثال هلن نیسنبام (۲۰۰۴) پس از نقد به نظارت عمومی در ایالات متحده معتقد است این امر حق حفظ حریم خصوصی را نقض می‌کند. او استدلال می‌کند که برای حفاظت کافی از حریم خصوصی در خصوص نظارت عمومی، باید به هنجارهای زمینه‌ای خاصی توجه کرد که تنها مستلزم جمع‌آوری و انتشارات اطلاعات متناسب با زمینه مورد نظر باشد.

۲. به‌عنوان مثال مارتین بلاو (۲۰۱۳) استدلال می‌کند که حریم خصوصی باید در قالب معرفتی درک شود. او معنای داشتن (درجاتی از) حریم خصوصی را این می‌داند که دیگران روابط معرفتی مهمی در موضوعاتی که فرد مایل به محرمانه نگاه داشتن آن‌ها است ندارند. وی حریم خصوصی را یک رابطه سه‌گانه میان یک سوژه (S)، مجموعه‌ای از گزاره‌ها (P) و افراد/ فرد دیگر (O) می‌داند. به‌عقیده وی، S زمانی از حریم خصوصی کامل برخوردار خواهد بود که O دسترسی معرفتی به P نداشته باشد. در مقابل اگر O بیشترین دسترسی معرفتی را با P داشته باشد S حریم خصوصی ندارد.

۳. Data protection

۴. Data security

۵. Privacy by design

۶. Data minimalization

۷. Access control

در میان انواع رویکردهایی که به آن‌ها اشـاره شـد، بیشـتر پژوهش‌گران اخلاق هوش مصنوعی، رویکرد خوشه‌ای از عناصری مانند «رضایت آگاهانه» را برای مطالعه یادگیری ماشـین اتخاذ می‌کنند[۱]. عقیده آن‌ها بر این اسـت که این رویکرد، کمک می‌کند که یک عامل انسانی قادر به تصمیم‌گیری و عمل موجه در برابر تصمیمات و خروجی‌های ماشین باشـد (Muller, 2021: 18). لذا ما نیز به بررسـی رویکرد خوشـه‌ای و عنصـر رضـایت آگاهانه در مباحث آتی تمرکز خواهیم کرد.

٤. کدری یادگیری ماشین در برابر حریم خصوصی

دومین مبحث اخلاقی پژوهش حاضر، کدری/شفافیت است. با تولید وسیع داده‌ها و اهمیت یافتن «اخلاقِ داده‌ها[۲]» و «اخلاقِ کلان داده‌ها[۳]» توجه فیلسوفان به موضوع شفافیت/کدری در اخلاق هوش مصـنوعی جلب شـد (Muller, 2021a). شـفافیت/کدری را می‌توان در معنای گوناگون و ابعاد مختلفی موردبررسـی قرار می‌گیرد. در اینجا مقصـود از شـفافیت مشـخص بودن روند رسـیدن ماشـین به یک نتیجه برای انسـان و در مقابل کدری، عدم مشـخص بودن این روند اسـت[۴]. در یادگیری ماشـین ما با طیفی از کدری/شـفافیت روبرو هستیم و می‌توان سه نوع کدری را از یکدیگر تمیز داد: کدریِ کم‌عمق[۵]، کدریِ استاندارد[۶] و کدریِ عمیق[۷] (Muller, 2021:18). با توجه تفاوت اشکالِ مختلف کدری در یادگیری ماشین، برای بررسی دقیق‌تر لازم خواهد بود به‌صورت مجزا به آن‌ها بپردازیم.

٤-١. کدری کم‌عمق

در سیستم‌های تصمیم‌گیری خودکار و سیستم‌های پشتیبان که از هوش مصنوعی استفاده

---

۱. برای مشاهده نمونه این پژوهش‌ها رجوع کنید به (Muller, 2021).

۲. Data ethics

۳. Big data ethics

۴. به‌عنوان مثال مقصود ما شفافیت در نحوه جمع‌آوری داده‌ها، برچسب‌گذاری، بررسی از جهت وجود سوگیری (Bias) در آن‌ها و موارد مشابه دیگر نبوده و شایسته است در پژوهش‌های دیگر مورد توجه پژوهشگران قرار گیرند.

۵. Shallow opacity

۶. Standard opacity

۷. Deep opacity

می‌کنند، افرادی که به سیستم دسترسی دارند (متخصصان)، ممکن است قادر باشند از علت و روند رسیدن به تصمیمات مطلع شوند و اصطلاحاً سیستم برای آن‌ها «شفاف» باشد. در طرف مقابل افرادی که سیستم از داده‌های آن‌ها استفاده می‌کند (غیرمتخصص)، از دلیل و شیوه تصمیم‌گیری نامطلع بوده و سیستم برای آن‌ها «کدر» است. گزارش برخی پژوهش‌ها حاکی از آن است که این سیستم‌ها بخشی از «ساختارهای قدرت[۱]» هستند (ibid:19).

برای روشن شدن این سطح از کدری ذکر یک مثال مفید خواهد بود. فرض کنید یک فرد نیازمند پیوند عضو (R) و نفر اول در لیست انتظار دریافت عضو است. فردی دیگر (G) می‌خواهد عضو خود را اهدا کند. احتمالاً عضو باید به فرد اول لیست انتظار، یعنی R برسد. در سیستم‌های هوش مصنوعی داده‌های افراد اهدا کننده و دریافت کننده با جزئیات مختلف ثبت می‌شوند. در این شرایط R متوجه می‌شود که فردی قصد اهدای عضو دارد، اما R هیچ اطلاعاتِ شخصی و حتی نام G را نمی‌داند و اهدا به‌صورت ناشناس صورت می‌گیرد. سیستم (M) با توجه به بررسی داده‌ها به این نتیجه می‌رسد که عضو اهدایی نباید به نفر اول لیست انتظار برسد (C)، چراکه برخی داده‌های پرونده پزشکی R نشان می‌دهد عضو اهدایی با توجه به شرایط او مناسب نیست. در این شرایط فرد دیگری که داده‌های پرونده او با عضو اهدایی تناسب دارد انتخاب شده و عضو به او پیوند می‌شود. با اینکه M از روی داده‌هایی R تصمیم‌گیری کرده است، علت این تصمیم‌گیری برای R نامشخص بوده و به‌عبارتی کدر است. حال متخصصینی (S) که به سیستم دسترسی دارند، ممکن است بتوانند از علت تصمیم‌گیری و داده‌های R و G مطلع شوند.

به‌صورت خلاصه:

۱- M از داده‌های R استفاده می‌کند و به نتیجه C می‌رسد.

---

۱. Power Structure: در اینجا ساختارهای قدرت به‌معنای ایجاد فرآیندهایی است که فرصت را برای مشارکت انسانی در تصمیم‌گیری محدود کرده و شرایط این محدودیت را به آنها تحمیل می‌کنند که این موضوع بیشتر در مباحث سیاست هوش مصنوعی مورد توجه قرار می‌گیرد (Danaher, 2016: 245). در این سیستم‌ها معمولا امکان مشارکت در تصمیم‌گیری برای افرادی که سیستم از داده‌های آن‌ها استفاده می‌کند محدود می‌شود. این سیستم‌ها با ساختارهای مشخصی از دسترسی طراحی می‌شوند. افرادی که سیستم از داده‌های آن‌ها استفاده می‌کند در اینکه این ساختارها چگونه طراحی شوند یا دسترسی‌ها تغییر کنند معمولا انتخابی ندارند (ر.ج Muller, 2021؛ Muller, 2021؛ Diakopoulos 2015).

۲- R از C مطلع می‌شود اما از علت رسیدن M به C اطلاعی ندارد و درنتیجه M برای R کدر است.

۳- S از C مطلع می شود و می‌تواند مطلع شود چگونه M به C رسیده است و درنتیجه M برای S شفاف است.

٤- داده‌های R و G ممکن است در دسترس S قرار بگیرد.

همان‌طور که در ابتدای این بخش مطرح شد چنین سیستمی برای فردی که سیستم از داده‌های او استفاده می‌کند (R) کدر است. در این حالت، سیستم برای فردی که به سیستم دسترسی دارد (S) شفاف است. در چنین سیستمی ساختارهای قدرت هستند که رضایت آگاهانه و شفافیت را تعیین می‌کنند. به‌عبارتی‌دیگر، ساختار قدرت تعیین می‌کنند که سیستم برای M شفاف باشد یا خیر و حریم خصوصی او حفظ شود یا خیر. در این سطح، محدودیت‌ها یا تأثیرات ذاتی هوش مصنوعی ما را با چالش‌های اخلاقی روبرو نکرده بلکه طراحی و سیاست‌گذاری تعیین‌کننده است. پس به صورت خلاصه می‌توان این‌طور گفت که در شفافیت کم‌عمق، این هوش مصنوعی نیست که مسئله شفافیت و حفاظت از حریم خصوصی را پدید می‌آورد و این موضوعات به ساختارهای قدرت بازمی‌گردند.

۲-٤. کدری استاندارد یا جعبه سیاه

برخی از تکنیک‌های یادگیری ماشین، سیستم‌ها مشابه شبکه‌های عصبی[1] طراحی می‌شوند.

---

۱ **Artificial Neural Networks (ANN): این شبکه‌ها از چند لایه تشکیل می‌شوند؛ یک لایه ورودی** (Input Layer)، یک لایه خروجی (Output Layer) و یک یا چند لایه پنهان (Hidden Layer) که میان لایه ورودی و خروجی قرار می‌گیرند. زمانی که بیش از یک لایه میانی وجود داشته باشد به آن شبکه عصبی عمیق (Deep Neural Network) و به روش یادگیری ماشین، یادگیری عمیق (Deep Learning) گفته می‌شود. ارتباط داده‌ها بین لایه‌ها ممکن است از یک جهت باشد که از این روش در سیستم‌های پیش‌خور (Feed-forward) استفاده می‌شود و یا ممکن است این ارتباط در جهت‌های مختلف صورت بگیرد که از آن در سیستم‌های بازگشتی (Recurrent) استفاده می‌شود. این شبکه‌های عصبی ماشین از طریق یک سیستم بازخورد، خروجی‌ها را براساس داده‌های ورودی تغییر می‌دهند تا ماشین یاد بگیرد (Worden et al., 2023).

یادگیری این شبکه‌ها به سه شیوه تقسیم‌بندی می شوند[۱]: یادگیری نظارتی[۲]، نیمه‌نظارتی[۳] و غیرنظارتی[۴] (Muller, 2021: 19). البته در بسیاری از سیستم‌ها، این شیوه‌ها در کنار

---

۱. در برخی از شیوه یادگیری در چهار دسته یادگیری نظارتی، غیرنظارتی، نیمه نظارتی و تقویتی (Reinforcement) طبقه‌بندی (Bhatnagar, 2019: 276) و تحت عنوان پارادایم (Paradigms) یادگیری ماشین نامیده می‌شوند. البته برخی نیز تعداد پارادایم‌ها را تا ۱۰ مورد قابل تقسیم‌بندی می‌دانند (Emmert-Streib & Dehmer: 2022). قابل توجه است که بیشتر مواقع روش نیمه‌نظارتی و تقویتی را در یک دسته قرار داده و سه پارادایم اصلی برای یادگیری ماشین در نظر گرفته می‌شود (Muller, 2021: 19) که ما نیز همین رویکرد را دنبال می کنیم.

۲. Supervised: در شیوه یادگیری نظارتی، داده‌های آموزشی برچسب‌گذاری می‌شوند. به‌عبارتی در این شیوه، هر داده آموزشی برچسبی دارد که آن را به یک خروجی مشخص مرتبط می کند (Kalita, 2022: 4). برچسب‌گذاری معمولا توسط عوامل انسانی صورت می گیرد. در این شیوه یادگیری، ماشین به‌دنبال یافتن فرضیه (مدلی) با تقلید از داده‌های آموزشی است. در حقیقت در مسیر یادگیری برچسب‌های داده‌های آموزشی، به نوعی راهنمای ماشین، برای یافتن فرضیه هستند. آن‌ها به ماشین نشان می‌دهند که کدام داده مربوط به کدام خروجی است (Jung, 2022: 12-13). به‌صورت کلی، دو نوع یادگیری نظارتی وجود دارد: طبقه‌بندی (Classification) و رگرسیون (Regression). زمانی که مجموعه داده‌ها، شامل داده‌هایی است که از یکدیگر گسسته هستند، از شیوه یادگیری نظارتی و الگوریتم‌های طبقه‌بندی استفاده می‌شود. زمانی که مجموعه داده‌های ما، داده‌های عددی باشد از الگوریتم‌های رگرسیون استفاده می‌شود. با عنایت به اینکه در شیوه این شیوه برچسب‌گذاری توسط عوامل بیرونی، انجام می‌شود این روش را نظارتی می‌نامیم (Kalita, 2022: 5-6).

۳. Semi-supervised: در خصوص یادگیری تقویتی یا نیمه‌نظارتی دو دیدگاه مختلف وجود دارد. برخی عقیده دارند که یادگیری تقویتی، شباهتی به یادگیری نظارتی و غیرنظاری ندارد؛ و در مقابل برخی معتقدند یادگیری تقویتی برخی از ویژگی‌های هر دو شیوه را دارد (Kalita, 2022: 7). یادگیری تقویتی (RL) تا حدودی مشابه شرطی‌سازی ابزاری (instrumental conditioning) در فیزیولوژی است. در شرطی‌سازی ابزاری، حیوانات ارتباط بین محرک و پاسخ را یاد می‌گیرند؛ به‌گونه‌ای که با توجه به یک محرک یا حالت محیطی، حیوان پاسخ یا عملی را امتحان می کند. اگر نتیجه پاسخ برای یک محرک مشخص مثبت باشد، ارتباط بین محرک و پاسخ تقویت می‌شود. در حقیقت ایده پارادایم یادگیری تقویتی این است که ماشین از طریق فرآیندی مشابه آزمون و خطا یاد بگیرد (Rolf et al., 2022: 4).

۴. Unsupervised: در یادگیری ماشین، گاهی نیاز نیست که از برچسب‌گذاری برای راهنمایی ماشین استفاده شود. در این شیوه، یادگیری یک فرضیه (مدل)، براساس ساختار ذاتی داده‌ها صورت می‌گیرد (Jung, 2022: 13). در حقیقت با توجه به اینکه برچسب گذاری در این روش استفاده نمی‌شود، یادگیری بدون عامل خارجی ناظر انجام می‌شود و به همین دلیل به آن یادگیری غیرنظاری گفته می‌شود (Kalita, 2022: 5-6). یادگیری غیرنظارتی شامل متدهای مختلفی است که میان آن‌ها می‌توان با متد خوشه‌بندی و روش‌های یادگیری ویژگی اشاره کرد (Jung, 2022: 13). در یادگیری غیرنظاریتی داده‌های ورودی بسیار بااهمیت خواهند بود؛ چراکه مثلا اگر داده‌ها دارای سوگیری باشند،

یکدیگر به‌کار می‌روند[۱] (Emmert-Streib & Dehmer, 2022).

در شیوه یادگیری نظارتی، ناظر/ناظران به ماشین می‌گویند آیا خروجی/پیش‌بینی ماشین درست است یا خیر. در روش نیمه‌نظارتی ناظر/ناظران به ماشین نمی‌گویند کدام خروجی/پیش‌بینی صحیح یا غلط است، بلکه اهداف اصلی‌تر به آن گفته می‌شود به‌عنوان‌مثال از ماشین رسیدن به نتیجه به‌جای شیوه رسیدن به نتیجه خواسته می‌شود. در روش غیرنظارتی، هیچ بازخوردی به ماشین داده نمی‌شود که آیا یک الگو درست است یا خیر و ماشین بر اساس ساختار ذاتی داده‌ها به یک مدل نهایی می‌رسد. با استفاده از این تکنیک‌ها ماشین قادر به یافتن الگوها از داده‌ها و رسیدن به مدل خواهد بود. متخصصین ماشین‌هایی که از روش غیر نظارتی استفاده می‌کنند، اگرچه الگوها و داده‌ها را مشاهده می‌کنند، اما نمی‌دانند چطور ماشین این الگوها را به دست آورده است. در این شیوه یادگیری، خروجی برای فرد غیرمتخصص و متخصصین «کدر» خواهد بود[۲]. البته باید توجه داشت میزان کدر بودن برای متخصص و غیرمتخصص در یک سطح نیست و یک متخصص ممکن است پس از بررسی بتواند علت تصمیم‌گیری ماشین را تا حدودی بیابد (Muller, 2021: 19). اگر این موضوع را با کدری کم‌عمق مقایسه کنیم مشاهده می‌شود که تفاوت در این است که در کدریِ کم‌عمق، روند رسیدن به خروجی ممکن است برای افراد غیرمتخصص «کدر» و برای متخصص «شفاف» باشد؛ اما در کدری استاندارد روند رسیدن به خروجی برای متخصص و غیرمتخصص با درجات مختلفی کدر است.

برای برطرف شدن کدری استاندارد در سیستم‌هایی که از روش غیرنظارتی استفاده می‌کنند، می‌توان از هوش مصنوعی درون‌نما[۳]، استفاده کرد و شفافیت را افزایش داد. هدف از هوش مصنوعی درون‌نما به‌وجود آوردن امکان درک عوامل چرایی تصمیم‌گیری‌های

یادگیری و مدل دارای سوگیری خواهند بود. اگرچه مبحث سوگیری در موضوع شفافیت دارای اهمیت است، اما با توجه به اینکه در راستای اهداف و پرسش‌های پژوهش حاضر نیست در اینجا بیشتر مورد نظر قرار نگرفته است.

۱. به‌عنوان مثال در پروژه گوگل اسکالر از پارادایم‌های گوناگون در کنار یکدیگر استفاده می‌شود (Emmert-Streib & Dehmer, 2022).

۲. هرچه سیستم با نظارت بیشتری طراحی شود، شفافیت افزایش پیدا خواهد کرد. با اینحال برای حل مسائل گوناگون نمی‌توان تنها از روش نظارتی استفاده کرد و چنین سیستم‌هایی قادر به شناخت الگوهای کشف نشده برای ما نیستند؛ لذا انتخاب میزان نظارت همواره امری دلخواهانه نیست.

۳ Explainable Artificial Intelligence (XAI)

ماشین است. با توجه به روش‌های گوناگون یادگیری ماشین و ترکیب این روش‌ها برای مقاصد گوناگون، در هوش مصنوعی درون‌نما نیز از روش‌های مختلفی استفاده می‌شود[۱] (Holzinger et al., 2022). اما به‌صورت کلی، سیاست‌گذاری برای استفاده از روش‌های هوش مصنوعی درون‌نما می‌تواند شفافیت را برای فرد متخصص و غیرمتخصص افزایش دهد. تحقیقات گسترده‌ای که در این زمینه انجام‌شده نشان می‌دهد این روش می‌تواند تا حدود زیادی مسئله کدری استاندارد را برطرف کند (Longo et al., 2020; Zednik & Boelsen, 2022; Zednik, 2021). لذا درمجموع موارد دو بخش کدری کم‌عمق و کدری استاندارد می‌توان نتیجه گرفت که چالش شفافیت/کدری در سطح کدری کم‌عمق با راه‌کار سیاست‌گذاری و در سطح کدری استاندارد با راه‌کار سیاست‌گذاری و فنی قابل‌رفع باشد.

۳-٤. کدری عمیق

حالت سومی نیز وجود دارد که در آن نمی‌توان به‌سادگی به شفافیت دست یافت. برای مشخص شدن این موضوع باید به استفاده از تحلیل داده‌ها[۲] در هوش مصنوعی توجه کرد. اگر مجموعه‌ای از داده‌ها داشته باشیم، پیش از تحلیل نمی‌دانیم نتیجه تحلیل داده‌ها چه خواهد بود. ممکن است تصور شود داده‌ها حاوی اطلاعات نیستند. اما زمانی که با انبوهی از داده‌ها روبرو هستیم الگوها و ارتباط داده‌ها می‌تواند اطلاعات گوناگونی را از آن‌ها استخراج کند؛ البته مشخص نیست چه میزان اطلاعات در انبوه داده‌ها وجود دارد. در این شرایط می‌توان به الگوها و اطلاعاتی دست یافت که پیش‌تر بر اساس داده‌های اولیه از وجود آن‌ها ناآگاه بوده‌ایم. به چنین شرایطی که نمی‌دانیم و نمی‌توانیم بدانیم انبوه داده‌ها حاوی چه اطلاعاتی هستند، کدری عمیق گفته می‌شود (Muller, 2021: 20). با این صورت‌بندی از شکل سوم کدری، متوجه می شویم در ابعاد داده‌های انبوه و با استفاده از تحلیل داده‌ها در یادگیری ماشین مسئله حریم خصوصی و شفافیت همچنان حل‌نشده باقی

---

۱. از جمله مهم‌ترین روش‌های هوش مصنوعی درون‌نما می‌توان به روش‌های LIME، Anchors، GraphLIME، LPR، DTD، TCAV، XGNN، SHAP، ASV، Break-Down، Shapley Flow اشاره کرد (Holzinger et al., 2022).

۲. Data analytics

می‌ماند و به‌سختی ممکن است قابل برطرف کردن باشد. به عبارتی ما نمی‌دانیم: ۱- چه اطلاعاتی ممکن است به دست آید، لذا نمی‌توانیم اطمینان داشته باشیم که حریم خصوصی افرادی که از داده‌های آن‌ها استفاده می‌شود تا چه اندازه در خطر خواهد افتاد و ۲- الگوهایی جدید به دست خواهند آمد که پیش‌تر حتی از وجود آن‌ها مطلع نبوده‌ایم و با کدری عمیق روبرو هستیم.

۵. بحث و بررسی

همان‌طور که در ابتدا مطرح شد، اهداف اصلی این پژوهش بررسی ارتباط میان شفافیت و حریم خصوصی در هوش مصنوعی برای نقد رویکردهایی است که راهکار برطرف شدن چالش‌های اخلاقی را در طراحی و اصلاحات فنی می‌دانند. لذا در این بخش در ابتدا به ارتباط میان شفافیت و حریم خصوصی پرداخته می‌شود و سپس راهکار متمرکز بر طراحی به نقد گذاشته می‌شود.

۱-۵. ارتباط شفافیت و حریم خصوصی

بر اساس آنچه موردبررسی قرار گرفت، در کدری کم‌عمق، ساختارهای قدرت؛ در کدری استاندارد، روش پردازش؛ و در کدری عمیق، محتوا و تنوع داده‌ها و توانایی ماشین در استخراج اطلاعات از آن‌ها، موضوع حفظ حریم خصوصی و کدری است.

مطابق مباحث بالا در کدری کم‌عمق، حفاظت از حریم خصوصی و رسیدن به شفافیت با اصلاح ساختارهای قدرت (سیاست‌گذاری) میسر بوده و به راه‌کارهای فنی نیاز نخواهیم داشت. در کدری استاندارد، نیاز به تصمیم‌گیری (به‌صورت سیاست‌گذاری برای استفاده از روش‌های درون‌نما) و اصلاحات فنی (به‌کارگیری روش‌های درون‌نما) خواهیم داشت. در کدری عمیق نمی‌دانیم داده‌ها حاوی چه اطلاعاتی هستند و چه الگوهایی از آن‌ها به دست خواهد آمد و لذا حفاظت از حریم خصوصی و دستیابی به شفافیت بسیار دشوارتر خواهد بود. از مجموع این موارد ممکن است به نظر برسد میان افزایش شفافیت و حفاظت بیشتر از حریم خصوصی یک همگرایی وجود دارد چراکه در سیستم‌هایی که کدری بیشتر است، حفاظت از حریم خصوصی دشوارتر به نظر می‌رسد. این نتیجه‌گیری اولیه را پیش‌فرض رویکردهایی است که در این پژوهش قصد نقد به آن‌ها را داریم، اما آیا شفافیت حداکثری مطلوب‌ترین حالت خواهد بود؟ این موضوع، از دو جهت قابل‌تأمل است:

۱- باید اشاره کرد که افزایش حفاظت از داده‌ها در هوش مصنوعی تنها محدود به موضوع شفافیت نیست و نباید حفظ حریم خصوصی را ذیل شفافیت بررسی کرد (یا برعکس). به‌عنوان‌مثال برخی از ابزارهای امنیتی این امکان را پدید می‌آورند که بتوان بدون شفافیت، محافظت بیشتری از داده‌ها انجام داد (Hagendorff, 2023). لذا بدون افزایش شفافیت نیز در برخی موارد می‌توان حفاظت از حریم خصوصی را بهبود بخشید و برای حفاظت از حریم خصوصی روش‌ها و راهکارهای دیگری نیز وجود دارند.

۲- افزایش شفافیت، خود به دور از مشکلات نیست. گاهی افزایش شفافیت می‌تواند به قیمت کاهش امنیت و به طبع آن در خطر افتادن حریم خصوصی تمام شود؛ به‌عنوان‌مثال احتمال هک شدن سیستم افزایش پیدا می‌کند (Whittlestone et al., 2019). به‌عبارتی‌دیگر، افزایش شفافیت گاهی می‌تواند مشکلات ثانویه ایجاد کند.

از مجموع دو نکته یاد شده، به چهار نتیجه دست می‌یابیم:

۱- برخلاف تصور اولیه، یک همگرایی واقعی میان شفافیت و حفظ حریم خصوصی وجود ندارد و گاهی شفافیت منجر به کاهش حفاظت از حریم خصوصی می‌شود.

۲- راهکارهای دیگری برای حفظ حریم خصوصی خارج از مبحث شفافیت وجود دارد.

۳- شفافیت همواره مطلوب نیست.

۴- گاهی برطرف کردن برخی از چالش های اخلاقی خود می‌توانند منجر به تقویت چالش‌های اخلاقی دیگر در یادگیری ماشین شوند.

تا به اینجا مشخص می‌شود که این پیش‌فرض که چالش‌های اخلاقی در هوش مصنوعی همگرا هستند، تصویر درستی نیست و می‌توان آن را با نمونه شفافیت و حریم خصوصی نقض کرد.

۲-۵. نقد به طراحی اخلاقی هوش مصنوعی

در ادامه با مطرح کردن سه چالش کمی‌سازی، محدودیت‌های فنی و موضوعات فرااخلاقی به نقد رویکردهایی پرداخته می شود که راهکار حل مشکلات اخلاقی در هوش مصنوعی را به طراحی و اصلاحات فنی محدود می‌سازند. قابل‌توجه است که از منظر فلسفی، این رویکردها عموماً بر واقع‌گرایی اخلاقی تأکید دارند (Häggström, 2021) و اتخاذ چنین موضوعی خود به‌دور از مخالفت نخواهد بود.

مطابق با اولین چالش، به کارگیری قواعد و کدهای اخلاقی در هوش مصنوعی به زبان

ریاضیات انجام می‌شود؛ به عبارتی موضوعات اخلاق هوش مصنوعی را به معیارهای کمی برای رسیدن به اهداف و شرایط مشخص تبدیل می‌کنیم (Hagendroff, 2022). برای روشن شدن این چالش به موضوع رضایت آگاهانه فرد رجوع می‌کنیم. یک پرسش این است که آیا رضایت آگاهانه را می‌توان با معیار کمی تضمین کرد؟ برای رضایت آگاهانه، نیازمند آگاهی از روندی پیچیده در ماشین هستیم. اگر آگاهی را به دانستن هدف اصلی ماشین از داده‌ها تقلیل دهیم، این پاسخ چندان رضایت‌بخش نخواهد بود؛ چراکه در دل خود توضیح حریم خصوصی را بی‌معنا می‌کند. به عبارتی همان‌طور که در تعریف هوش مصنوعی ذکر شد، هوش مصنوعی نوعی سیستمِ محاسباتی است که برای رسیدن به هدف یا اهداف موردنظر رفتارهای هوشمند اتخاذ می‌کند. درنتیجه می‌توان گفت هر هوش مصنوعی رضایت آگاهانه را تضمین می‌کند چراکه برای هدفی مشخص طراحی‌شده و لذا با مسئله اخلاقی روبرو نیستیم.

اگر مقصود از رضایت آگاهانه اطلاع از روند رسیدن به نتیجه باشد، آن‌گاه نیازمند شفافیت خواهیم بود. در اینجا دو نکته قابل‌تأمل است. اول، مطابق بررسی‌های بالا دانستیم که شفافیت خود چالش‌هایی به همراه دارد (مثلاً می‌تواند گاهی حریم خصوصی را در خطر بیندازد)؛ و دوم برای کدری عمیق گاهی رسیدن به شفافیت امکان‌پذیر نیست. حال برای مواردی که امکان دست‌یابی به شفافیت وجود دارد، مثلاً در شیوه درون‌نما ماشین به‌گونه‌ای طراحی می شود که روند رسیدن به نتیجه را ساده‌سازی می‌کند. در برخی موضوعات، ساده‌سازی برای انتخاب آگاهانه می‌تواند کاربردی باشد. گاهی تنها نیاز است روند رسیدن به نتیجه به‌صورت یک دیاگرام ارائه شود، اما در مسائل پیچیده، ساده‌سازی خود تبدیل به چالش می‌شود. در چنین شرایطی با این پرسش مواجه می‌شویم که چه میزان ساده‌سازی لازم است تا یک فرد غیرمتخصص بتواند اطمینان حاصل کند که آگاهانه تصمیم‌گیری کرده است؟ ساده‌سازی ماشین را در قالب یک مدل می‌توان بررسی کرد. درک یک مدل پیشرفته می‌تواند برای یک متخصص هم چالش‌برانگیز باشد. به‌خصوص آنکه ممکن است ماشین الگوها و مدل‌های جدیدی از داده‌ها بیابد که پیش‌تر با آن‌ها آشنایی نداشته‌ایم. این موضوع زمانی دشوارتر می‌شود که می‌دانیم ماشین همواره در حال یادگیری است و مدل در طول زمان تغییر می‌کند. در این شرایط با یک پرسش معرفت‌شناسی قدیمی در فلسفه علم روبرو می‌شویم. اینکه کدام یک از مدل‌های ساده یا مدل‌های پیچیده فهم درست‌تری به دست دهند هنوز در میان فیلسوفان

علم موردبحث بوده و به‌اتفاق نظری دست نیافته‌ایم (به‌عنوان‌مثال ر.ج Bokulich, 2008; Kuorikoski & Ylikoski, 2015; Strevens, 2008) و در اینجا نیز این پر سش همچنان پابرجاست. همچنین قابل‌توجه است که برای موضوعاتی که نیازمند اطلاع کیفی هستیم، معیارهای کمی برای سنجش آگاهی از روند رسیدن ماشین به نتیجه چندان رضایت‌بخش نخواهد بود.

چالش دوم این است که به کارگیری کدهای اخلاقی در هوش مصنوعی گاهی با محدودیت‌های فنی روبرو است. اگرچه فیلسوفان و متخصصین اخلاق و حقوق، ممکن است بتوانند در رسیدن به تحلیل‌های مفهومی و تفاسیر اصطلاحی موفق عمل کنند، اما دست‌آخر آن‌ها کدهای اخلاقی را پیاده‌سازی نمی‌کنند و یک نظارت همیشگی بر تصمیمات سیستم ندارند. در عمل این طراحان هوش مصنوعی هستند که می‌توانند با درکی که از این کدها به دست آورده‌اند و با توجه به ظرفیت‌های فنی، کدها را در ماشین پیاده‌سازی کنند. از پس این محدودیت‌های فنی، گاهی مشکلات متعددی بروز پیدا می‌کند. به‌عنوان‌مثال زمانی که از هوش مصنوعی درون‌نما استفاده می شود قصد قابل‌فهم کردن تصمیم ماشین برای انسان است. با این حال اطلاع از این‌که یک خروجی سیستم هوش مصنوعی در چه فرآیندی به‌دست آمده، به معنای موجه دانستن آن خروجی نیست. به سخنی دیگر، اینکه سیستم بر چه اساس تصمیم‌گیری کرده است (درون‌نمایی)، توجیه‌کننده آن تصمیم نیست (ibid). حال حتی اگر موجه بودن برای تمامی اشکال کدری نیز میسر شود، کماکان پیاده‌سازی کدهای اخلاقی بر اساس ظرفیت‌های فنی پاسخ رضایت‌بخشی نخواهد بود. می‌توان این ایراد را با یک پرسش بیان کرد: اگر ظرفیت‌های فنی امکان پیاده‌سازی برخی از کدهای اخلاقی را نداشته باشند، مجاز هستیم از آن کدهای اخلاقی چشم‌پوشی کنیم و یا شایسته است از روش‌های دیگر نیز بهره ببریم؟

چالش سوم از محدودیت‌های طراحی اخلاقی هوش مصنوعی است. سه شکل کلی از طراحی اخلاقی هوش مصنوعی وجود دارد. در شکل اول هوش مصنوعی به‌گونه‌ای طراحی می‌شود که تصمیمات اخلاقی، تحت کنترل و توسط افراد متخصص گرفته می شود. این نوع طراحی زمانی قابل‌استفاده است که ماشینی عملکردی شفاف، روند آن ساده است، تصمیم‌گیری محدود بوده و عامل انسانی می‌تواند نظارت و تصمیم‌گیری به‌موقع داشته باشد. در شکل دوم، ماشین ظرفیت ارزیابی و پاسخگویی به چالش‌های

اخلاقی در موضـــوع مشـــخص را دارد. به‌عنوان‌مثال تمامی شـــرایط از قبل در طراحی پیش‌بینی شده و ماشین بر اساس عملکرد، پاسخ اخلاقی مشخصی را یافته و طبق آن عمل می‌کند. در شکل سوم، ماشین یک عامل اخلاقی[۱] دانسته می شود. در این حالت ماشین به‌صـورت خودکار خروجی‌ها و تصـمیمات اخلاقی را برعهده گرفته و یادگیری او شـامل یادگیری اخلاق نیز می شود (Wallach & Allen, 2009). پروا ضح است که دو حالت اول را نمی‌توان برای کلان داده‌ها که با تصمیم‌گیری‌های فراوان و روندهای پیچیده و شاید ناشناخته روبرو هستیم به کار برد.

در شکل سوم طراحی اخلاقی ماشین دو رویکرد عمده وجود دارد. در رویکرد نخست، قضاوت‌های توصیفی[۲] اخلاقی وابسته به شرایط گوناگون به‌عنوان داده‌های آموزشیِ ماشین به کار گرفته می شـــوند. در رویکرد دوم قواعد تجویزی[۳] اخلاقی، داده های آموزشـــی و خروجی‌های ماشـــین را هدایت می‌کنند؛ به این صـــورت که اگر قضـــاوت‌های انســـانی و داده های آموزشـــی، هنجارهای اخلاقی را نقض کنند، ماشـــین آن‌ها را بر اســـاس قواعد تجویزی که در ابتدا دریافت کرده، بازنویسی و اصلاح می‌کند. این دو رویکرد معمولاً در کنار هم به کار می‌روند. در این شـــکل ماشـــین از قضـــاوت‌های فردی به‌عنوان داده‌های آموزشـــی اســـتفاده می‌کند و بخشـــی از یادگیری ماشـــین، یادگیریِ یافتنِ الگوهایِ اخلاقی اســـت و ســـپس با قواعد تجویزی، داده‌ها و قضـــاوت‌هایی که هنجارهای اخلاقی را نقض می‌کنند اصلاح می شوند. اگرچه شکل سوم طراحی اخلاقی ماشین برای برخی از مسائل اخلاقی، عملکرد خوبی از خود نشـــان داده، اما همچنان برخی مشـــکلات وجود دارد[۴]. به‌عنوان‌مثال در موضـــوعات فراخلاقی که اتفاق‌نظر وجود ندارد، طراحی با مشـــکل روبرو است. همچنین قابل‌ذکر است که این شکل سوم رو به سوی گذشته دارد؛ به عبارتی هوش مصـــنوعی مشـــکلات اخلاقی که در آینده با آن‌ها روبرو می شـــویم را باید با توجه به آموزش‌ها و کدهای اخلاقی که بر اساس شرایط گذشته ارائه شده‌اند، موردبرر سی قرار

---

۱. Moral Agent

۲. Descriptive

۳. Prescriptive

۴. به‌عنوان موفق‌ترین نمونه می‌توان به Delphi اشاره کرد که البته این هوش مصنوعی نیز با برخی چالش‌های اخلاقی و فرااخلاقی روبرو است (Hagendorff & Danks, 2022).

دهد. حال آنکه عامل اخلاقی نمی‌تواند تنها با الگوهای گذشته رجوع کرده و گاهی نیاز به اتخاذ رویکرد هنجاری داریم (Hagendorff & Danks, 2022). لذا راهکار طراحی در موضوعاتی مانند حریم خصوصی با نگاه رو به گذشته و رویکرد توصیفی در اخلاق مواجه خواهد بود.

پیش‌تر اشاره شد که در کلان داده‌ها ممکن است الگوهای ناشناخته به دست آید؛ حال اگر در برابر الگوهای ناشناخته تنها به طراحی بسنده کرده و پیرو آن به الگوهای قدیمی رجوع کنیم، نمی توان تصمیم گیری های اخلاقی در الگوهای جدید را توجیه کرد. در این شرایط نمی‌دانیم تأثیر این الگوها چیست، چه شرایطی به وجود می آورند و آیا با رجوع به شرایط گذشته استفاده از این تصمیم‌گیری‌ها در شرایط جدید نیز توجیه‌پذیر خواهد بود یا خیر.

از مجموع مواردی که در بالا مطرح شد، می‌توان نتیجه گرفت که اگر تنها به طراحی هوش مصنوعی تکیه کنیم، محدودیت های گوناگونی وجود خواهد داشت که منجر می شود دست ما در برابر چالش‌های اخلاقی در برخی موقعیت‌ها خالی باشد. پیشرفت‌های شگرفتی در حوزه هوش مصنوعی رخ داده، اما اگر چشم بر روی محدودیت‌های فعلی ببندیم و توانایی های کنونی را بیش از آنچه هست بدانیم، دچار اشتباهات محاسباتی خواهیم شد. البته دو نکته در راستای انتقاداتی که مطرح شده وجود دارد. اول، اگرچه چالش‌های مختلفی در طراحی پیش روی ما قرار دارد، اما به این معنا نیست که مقصود این است که از ظرفیت هوش مصنوعی بهره برده نشود؛ دوم آنکه برای فائق آمدن بر چالش‌های موجود می‌توان از روش های دیگر در کنار تلاش برای غلبه بر محدودیت های طراحی و فنی استفاده کرد. قانون‌گذاری، پیشرفت در علوم داده‌ها و روش‌های مختلف دیگر می‌تواند کمک کند تا بخشی از محدودیت‌های موجود را برطرف کرد[1].

---

۱. تأکید ما بر این است که با توجه به محدودیت‌های موجود، بایسته است که از روش‌های مختلف در کنار هم و نه به‌عنوان جایگزین بهره برد. در بسیاری موارد نیز این رویکرد مورد توجه قرار گرفته است. به‌عنوان نمونه می‌توان به قانون‌گذاری GPDR اشاره کرد. البته این راه‌کارها نیز خالی از ایراد نبوده و به همین دلیل پیشنهاد استفاده از چندین راه‌کار شده است. به‌عنوان مثال سیلویا لو نشان می‌دهد که افشای الگوریتم از طریق گزارش‌های پایداری و حفاظت از افشاگران در برابر شرکت‌هایی که هوش مصنوعی را به‌کار می گیرد، می‌تواند جایگزین بهتری از قوانین سخت‌گیرانه‌ای مانند GDPR باشد که در برابر حفاظت از داده‌ها چندان موفق نبوده‌اند (Lu, 2022: 2087-2088).

نتیجه‌گیری

با توجه به اهداف مطرح شده، یافته‌های این پژوهش به دو بخش تقسیم‌بندی می شود. در بخش اول، نتایج نشـان می‌دهد که در مباحث اخلاق هوش مصـنوعی، میان شـفافیت/کدری و حریم خصو صی یک همگرایی وجود ندارد. به‌عبارتی‌دیگر، افزایش شفافیت تضمین‌کننده حفاظت از حریم خصـوصـی نیسـت. قابل‌توجه اسـت، درصـورتی‌که این دو مبحث به‌صـورت همگرا موردنظر قرار گیرند، ثمره آن می‌تواند در خطر افتادن حریم خصـوصـی باشـد. گاهی ممکن است شفافیت و حفاظت از حریم خصو صی در یک هوش مصنوعی با یکدیگر بهبود بیابند اما این موضـوع یک قاعده کلی نبوده و نباید آن را به معنای تقویت یکی توسـط دیگری دانسـت. همچنین مشخص شد برای حفاظت از حریم خصو صی در هوش مصنوعی روش‌های متفاوتی وجود دارد و همواره نیازمند شفافیت نخواهیم بود. این نتایج به صورت ضمنی مؤید این ا ست که چرا مباحث اخلاقی هوش مصـنوعی را به‌عنوان مباحث جداگانه و غیرهمگرا بررسـی کنیم. این عدم همگرایی نشـان می‌دهد ما با مباحث مهم و نه یک مبحث اخلاقی و زیرمجموعه‌های آن مواجه هستیم. در بخش دوم نتایج پژوهش نشان می‌دهد برای پا سخ به چالش‌های اخلاقی، اگرچه در بسـیاری موارد راه‌کارها و اصـلاحات فنی و طراحی نتیجه‌بخش اسـت، اما در برخی شـرایط با محدودیت‌هایی مواجه اسـت و به‌تنهایی برای همه شـرایط راهگشـا نخواهد بود. همان‌طور که راه‌کارهای فلسـفی و اخلاقی که نگاهی به سـاختارهای مختلف هوش مصـنوعی ندارند، می‌توانند به تحلیل‌های کلی که در بسیاری موارد ارتباطی با ساختارهای هوش مصنوعی ندارد منتج شـود؛ اگر راه‌کارهای فنی را نیز بدون توجه به شـرایط مختلف به‌عنوان راه‌کارهای کلی در نظر بگیریم باز به شیوه‌ای دیگر دچار اشتباه خواهیم شد. در چنین شرایطی نیاز است از سـایر روش‌ها مانند قانون‌گذاری، پیشـرفت در علوم گوناگون در کنار طراحی اخلاقی اسـتفاده شود تا بتوان بخشی از محدودیت‌های موجود را برطرف ساخت.

تعارض منافع

تعارض منافع ندارد.

ORCID

Mohammad Ali Ashouri Kisomi https://orcid.org/0000-0001-5663-1993

منابع

Barhamgi, M. Bertino, E. (2022). Special Issue on Data Transparency—Uses Cases and Applications. ACM Journal of Data and Information Quality (JDIQ), 14(2), 1-3. https://doi.org/10.1145/3494455

Bhatnagar, R. (2019). Unleashing Machine Learning onto Big Data: Issues, Challenges and Trends. In Aboul Ella Hassanien (Eds.), *Machine learning paradigms: Theory and application* (pp. 271-286). Cham: Springer. https://doi.org/10.1007/978-3-030-02357-7_13

Blaauw, M. (2013). The Epistemic account of privacy. Episteme, 10(2), 167-177. doi:10.1017/epi.2013.12

Bokulich, A. (2008). Reexamining the Quantum-Classical Relation, Cambridge: Cambridge University Press.

DeCew, J. (2018). Privacy. Edward N. Zalta (Eds.). *The Stanford Encyclopedia of Philosophy*. Retrieved April 8, 2023, from https://plato.stanford.edu/archives/spr2018/entries/privacy

Descartes, R. (1955). The philosophical works of Descartes, Voume 1 (E.S, Haldane & G.R.T, Ross Trans.). New York: Dover Publications.

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, 29(3): 245–268. doi:10.1007/s13347-015-0211-1

Diakopoulos, N. (2015). Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. *Digital Journalism*, 3(3): 398–415. doi:10.1080/21670811.2014.976411

El Naqa, I. Murphy, M. (2015). What is machine learning? In El Naqa, I. Li, R & Murphy, J (Eds.), Machine Learning in Radiation Oncology. Springer International Publishing.

Emmert-Streib, F. Dehmer, M. D. (2022). Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5), e1470. https://doi.org/10.1002/widm.1470

Falsinski, M. (2016). Introduction to Artificial Intelligence. Springer.

Franzoni, V. (2023). From Black Box to Glass Box: Advancing Transparency in Artificial Intelligence Systems for Ethical and Trustworthy AI. In: Gervasi, O., et al. Computational Science and Its Applications – ICCSA 2023 Workshops. ICCSA 2023. Lecture Notes in Computer Science, vol 14107. Springer, Cham. https://doi.org/10.1007/978-3-031-37114-1_9

Hagendorff, T. (2023). Information Control and Trust in the Context of Digital Technologies. In Varieties of Cooperation: Mutually Making the Conditions of Mutual Making (pp. 189-201). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-39037-2_9

Hagendorff, T. (2022). Blind spots in AI ethics. *AI and Ethics*, *2*(4), 851-867. https://doi.org/10.1007/s43681-021-00122-8

Hagendorff, T. Danks, David. (2022). Ethical and methodological challenges in building morally informed AI systems. *AI Ethics*, 1-14. https://doi.org/10.1007/s43681-022-00188-y

Hagendroff, T. Wezel, K. (2020). 15 challenges for AI: or what AI (currently) can't do. *AI & Society*, 35, 355-365. https://doi.org/10.1007/s00146-019-00886-y

Häggström, O. (2021). AI, orthogonality and the Muller-Cannon instrumental vs general intelligence distinction. arXiv e-prints, arXiv-2109. https://doi.org/10.48550/arXiv.2109.07911

Holzinger, A. Sarnati, A. Molnar, C. Biecek, P. Samek, W. (2022). Explainable AiMethods – A Bried Overview. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ECML 2020, July 2018, Vienna, Austria, Revised and Extended Papers* (13-38). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_2

Jobin, A. Ienca, M. Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Jung, A. (2022). Machine Learning, The Basics. Springer Singapore.

Kalita, J. (2022). Machine Learning: Theory and Practice. Boca Raton: CRC Press.

Kuorikoski, J. Ylikoski P. (2015). External Representations and Scientific Understanding. *Synthese,* 192(12), pp. 3817–37.

Lee, J. Kang, H. Lee, Y. Choi, W. Eom, J. Deryabin, M. Lee, E. Lee, J. Yoo, D. Kim, Y. No, Jong-Seon. (2022). Privacy-preserving machine learning with fully homoorphic encryption for deep neural network. *IEEE Access*, 10, 30039-30054. doi: 10.1109/ACCESS.2022.3159694.

Liu, Z. Guo, J. Lam, K. Zhao, J. (2022). Effective dropout-resilient aggregation for privacy-preserving machine learning. *IEEE Transactions on information forensics and security*. 18, 1839-1854. doi: 10.1109/TIFS.2022.3163592.

Longo, L. Goebel, R. Lecue, F. Kieseberg, P. Holzinger, A. (2022). Explainable artificial intelligence: Concepts, Applications, research challenges and visions. In *Machine learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-Make 2020, Dublin,*

Ireland, August 25-28, 2020, Proceeding (pp. 1-16). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_1

Lu, S. (2022). Data privacy, human rights, and algorithmic opacity. *Cal. L. Rev*, 110, 2087-2147.

Matheson, D. (2008). A Distributive Reductionism About the Right to Privacy. *The Monist*, 91(1), 108-129. https://doi.org/10.5840/monist200891115

Morozov, E. (2013). To save everything, click here: The folly of technological solutionism. New York: PublicAffairs.

Muller, V. (2009). Would you mind being watched by machines? Privacy concerns in data mining. *AI & society*, 23(4), 529-544. https://doi.org/10.1007/s00146-007-0177-3

Muller, V. (2012). Introduction: Philosophy and Theory of Artificial Intelligence. *Minds, & Machines*, 22(2), 67-69. https://doi.org/10.1007/s11023-012-9278-y

Muller, V. (2021). Deep opacity undermines data protection and explainable artificial intelligence. In Overcoming opacity in machine learning.

Muller, V. (2021a). Ethics of Artificial Intelligence and Robotics. The Stanford Encyclopedia of Philosophy. Retrieved April 7, 2023, from https://plato.stanford.edu/entries/ethics-ai/

Muller, V. Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. *Fundamental issues of artificial intelligence*, 555-572. https://doi.org/10.1007/978-3-319-26485-1_33

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review Association*, 79(2-1), 119-157.

Rolf, B. Jackson, I. Müller, M. Lang, S. Reggelin, T. Ivanov, D. (2022). "A review on reinforcement learning algorithms and applications in supply chain management". *International Journal of Production Research*, 1-29.

Russel, S. Norvig, P. (2020). Artificial intelligence: A Modern Approach, Fourth Edition. London: Pearson.

Sengan, S. Khalaf, O. I. Sharma, D. K. Hamad, A. A. (2022). Secure and Privacy-Based IDS for healthcare systems on E-medical data using machine learning approach. *International Journal of Reliable and Quality E-healthcare (IJRQEH),* 11(3), 1-11. Doi: 10.4018/IJRQEH.289175

Strevens, M. (2008). Depth: An Account of Scientific Explanation. Cambridge: Harvard University Press.

Thomas, W. (2015). Algorithms: From Al-Khwarizmi to Turing and Beyond. In G. Sommaruga & T. Strahm (Eds.), *Turing's revolution: The impact of his ideas about computability*, 29-42.

Turing, A. (1950). Can machine think. *Mind*, 59(236), 433-460.

Van den Hoven, J. Blaauw, M. Pietres, W. Warnier, Martjn. (2020). *Privacy and Information Technology*. The Stanford Encyclopedia of Philosophy. Retrieved April 6, 2023, from https://plato.stanford.edu/entries/it-privacy

Wallach, W. Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. New York: Oxford University Press.

White, A. (2014). Digital media and society: transforming economics, politics and social practices. UK: Springer.

Whittlestone, J. Nyrup, R. Alexandrova, A. Cave, S. (2019, January). The role and limits of principles in AI ethics: towards a focus on tensions. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195-200).

Worden, K., Tsialiamanis, G., Cross, E. J. Rogers, T. J. (2023). Artificial neural networks. In Machine Learning in Modeling and Simulation: Methods and Applications (pp. 85-119). Cham: Springer International Publishing.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & technology*, 34(2), 265-288. https://doi.org/10.1007/s13347-019-00382-7

Zednik, C. Boelsen, H. (2022). Scientific exploration and explainable artificial intelligence. *Minds and Machines*, 32(1). 219-239. https://doi.org/10.1007/s11023-021-09583-6

استناد به این مقاله: عاشوری کیسمی، محمدعلی، همگرایی حریم خصوصی و شفافیت، محدودیت‌های طراحی هوش مصنوعی، حکمت و *فلسفه*، ۲۰ (۷۸)، ۴۵-۷۳.

DIO: 10.22054/wph.2024.75680.2183