Journal of Philosophical Investigations
Print ISSN: 2251-7960 Online ISSN: 2423-4419
Homepage: https://philosophy.tabrizu.ac.ir

On the Relation of Emotion and Moral Capacity in Artificial Intelligence Technologies

Zahra Zargar

Assistant Professor in Institute for Science and Technology Studies, Shahid Beheshti University, Tehran, Iran. E-mail: Z_zargar@sbu.ac.ir

Article Info

Article type:
Research Article

Article history:
Received 12 September 2024
Received in revised from 19 December 2024
Accepted 07 January 2024
Published online 21 March 2025

Keywords:
emotions, Emotions and Morality, Artificial Moral Agency, Artificial Intelligence Ethics, Philosophy of Artificial Intelligence.

ABSTRACT

The unpreceded abilities of AI technologies have led to the emergence of new ethical issues; among them is the possibility of the moral agency of AI artifacts. There are many questions around this subject, including what are the necessary and sufficient conditions of being a moral agent? How can we examine those conditions in artifacts? What levels and degrees of agency are possible for artifacts? And what level of moral agency is proper for allocating a certain task to AI artifacts? There are wide discussions about factors that figure in the moral capacities of AI artifacts, and emotions are one of the frequently referred factors. Emotions are directly or indirectly relevant for examining AI's moral status. In this paper, we focus on the relation between emotions and moral capacities in AI technologies. Our main question is whether emotions play a positive role in the improvement of moral capacities or a negative role. We extract and articulate four arguments in defense of the positive role of emotions in the enhancement of AI moral capacities, including arguments from *moral sensitivity*, *bounded rationality*, *risk assessment*, *and culpability*. Then, we present other four arguments in defense of the negative role of emotions in moral capacities including arguments *from emotional hijacking, deceptive emotions, anthropomorphism and dehumanization paradox, and moral deskilling*. Finally, we analyze the debate by clarifying the point of contest between the two mentioned camps and discuss serious challenges of designing emotional AI.

Cite this article: Zargar, Z. (2025). On the Relation of Emotion and Moral Capacity in Artificial Intelligence Technologies. *Journal of Philosophical Investigations*, 19 (50), 19-40. https://doi.org/10.22034/jpiut.2024.63626.3875

© The Author(s).
Publisher: University of Tabriz.

Extended Abstract

Introduction

Technical artifacts are widely assumed to have moral status. However, there are various views about the morality of machines and its nature. One important issue in discussion about the morality of machines is the moral capacity and moral status of AI technologies. Due to AI's capacities for learning and reasoning, there are serious debates about the plausibility of AI's moral agency and moral patients (Johnson and Noorman, 2014) (Behdadi and Munthe, 2020) (Nylhom, 2019) (Franklin and Graesser, 1996) (Powers, 2013) (Brey, 2014) (Bringsjord, 2007). One of the frequently discussed factors in these debates is emotion. While emotion is mentioned in a large part of the literature, researchers' attitudes split over assuming either a positive or a negative role for emotion in enhancing the moral capacity of AI. This is the main subject we are concerned with in this paper. I suggest four arguments defending the positive role of emotion, and four arguments defending the negative role of emotion. Finally, I analyze arguments based on some of their assumptions about emotion and discuss the significant moral challenges of designing emotional AI.

Arguments for the Positive Role of Emotion in AI's Moral Capacity

The following arguments consider different ways in which emotions foster morality.

Emotions and Moral Sensitivity: Some philosophers argue that emotions play a critical role in grasping the morally significant aspects of situations. Farisco et al. say that emotional engagement attracts our attention to what is morally important. Emotions function like a moral compass that enables us to see the needs of others (Farisco et al., 2020, 2419). Véliz similarly argues that sentience works like an internal moral lab that guides our actions. Having the experience of hurt, pain, and pleasure makes us able to imagine the effect of our actions on others (Véliz, 2021, 493). Moreover, emotions provoke sensitivity to moral exemplars. Some researchers emphasize the role of the emotion of *admiration* in recognizing moral exemplars, which is essential in moral motivation and behavior (Govindarajulu et al., 2019, 2 & 5).

Emotions and Bounded Rationality: In any moral situation, there is an unlimited range of evidence and arguments relevant to decision-making. But the time for decision and action is limited. Our bounded rationality helps us find a *satisfying* way in a proper time (Simon, 1967, 32). Some emotions are useful for deciding on complicated matters. According to the somatic theory of emotion, whenever the difference of options is not clear, and consequences are uncertain, the somatic marker makes an overall feeling in respect to the total effects of an action. This emotion makes it possible for the organism to limit deliberation and choose a way (Wallach and Allen, 2009, 147-149).

Emotions and Risk Assessment: Recognition and assessment of risk are morally important, especially for complicated actions like designing new artifacts. Roeser argues that *moral emotions* are products of self-conscious processes with high degree of reflectivity and narrativity. These emotions improve our perception of possible risks (Roeser, 2012,106-107).

Emotions and Culpability/Rewardability: Penalties and rewards are assumed to be essential, both in moral development of individuals and the emergence of morality in communities. For example, according to Kohlberg's theory, in the early stages of moral development, rewards and punishments make a primary sense of rightness and wrongness in children. The capacity for experiencing emotions is essential for the utility of reward and punishment. Thus, Véliz claims that having sentience is a necessary condition for moral agency (Véliz, 2021, 48).

Arguments for the Negative Role of Emotion in AI's Moral Capacity

The first following argument is about the relation of emotion and artificial moral agency, and the rest arguments are about artificial moral patience.

Emotional Hijacking: According to some moral philosophies (like Kantian ethics) and some empirical evidence, passions deviate us from reasoning and acting morally. Therefore, some philosophers think artifacts without emotional states could be moral agents far better than humankind due to being immune to *emotional hijacking* (Butkus, 2020). Artificial emotionless moral agent even might be moral guides for human and elevate our moral standards (Storrs Hall, 2007,353-354).

The Delusion of Emotion: Research about human comfortability in dealing with AI shows that AI's emotional expression encourages people to trust them and ascribe autonomy and agency to them; thereby promotes the quality of AI-human teamwork (Scheutz and Crowell, 2007, 5). But in the absence of real emotional experience, implanting emotional expressions in AI sounds tricky. Expressing emotions without experiencing it might enable AI to manipulate human's emotional responses in a way similar to what psychopaths do (Storrs Hall, 2007, 291-292).

Anthropomorphism and Dehumanization Paradox: Having emotional appearance, makes AI look humanoid for people, while AI's essential properties like being designed by humans and for serving human ends, dehumanize them (Cappuccio et al., 2019, 26). This condition makes AI-human relation very similar to master-slave relation, in which masters subjugate their fellows. Reviving this relation could be corruptive for humans and foster vices in their character.

Moral Deskilling: Emotional AI, exemplified in some chatbots, is going to compensate emotional needs which previously were satisfied in social human-human relationships. Since in these technologies AI behaves exactly according to the user's desire and tends to be substituted for human social relationships, engagement in these relationships decreases the opportunities of cultivating virtues through challenges of real social relationships (Vallor, 2024, 152).

Conclusion

In the above arguments, various aspects of emotionality in AI were discussed. Despite their opposite attitudes toward the relation of emotion and moral capacity, examining some of their crucial assumptions helps to find conformity among them. Emotions in humans have both functional and phenomenal aspects. However, in AI at best we can implement the functional aspect. In the above arguments when the positive role of emotions is defended (moral sensitivity, bounded rationality, risk assessment, culpability) mostly the phenomenal aspect is considered. But when the negative role of emotion is highlighted (the delusion of emotion, anthropomorphism and dehumanization paradox, and moral deskilling arguments; but not emotional hijacking) the functional aspect is at work. AI's emotion -contrary to human's- is merely functional. Then, regarding all mentioned arguments, we can conclude that the threats of emotional AI for its moral capacities is significantly more than its opportunities.

پژوهش‌های فلسفی
شاپا چاپی: ۷۹۶۰-۲۲۵۱ شاپا الکترونیکی: ۴۴۱۹-۲۴۲۳
Homepage: https://philosophy.tabrizu.ac.ir
Journal of Philosophical Investigations

رابطه عواطف و ظرفیت اخلاقی در فناوری‌های هوش مصنوعی

زهرا زرگر

استادیار، پژوهشکده مطالعات بنیادین علم و فناوری، دانشگاه شهید بهشتی، تهران، ایران. رایانامه: Z_zargar@sbu.ac.ir

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۳/۰۶/۳۱

تاریخ بازنگری: ۱۴۰۳/۰۹/۲۸

تاریخ پذیرش: ۱۴۰۳/۱۰/۱۸

تاریخ انتشار: ۱۴۰۴/۰۱/۰۱

کلیدواژه‌ها:
عواطف، عواطف و اخلاق، عاملیت اخلاقی مصنوعی، اخلاق هوش مصنوعی، فلسفه هوش مصنوعی.

چکیده

فناوری‌های هوش مصنوعی، در مقایسه با مصنوعات پیشین دارای قابلیت‌هایی جدید هستند که مسائل اخلاقی متفاوتی را مطرح می‌کند. ازجمله این مسائل، بحث درباره عاملیت اخلاقی هوش مصنوعی است. مسئله عاملیت اخلاقی مصنوعی، پرسش‌های نظری و عملیاتی مختلفی را مطرح می‌سازد؛ مانند این که شروط لازم و کافی برای عاملیت اخلاقی چیست، چطور می‌توان برقراری این شروط را در مصنوعات بررسی کرد، عاملیت اخلاقی دارای چه سطوح و درجاتی است و هر سطح از عاملیت اخلاقی برای واگذاری چه نوع وظایفی مناسب است. یکی از موضوعات پرتکرار در بحث حول عاملیت اخلاقی مصنوعی، عواطف است. فیلسوفان مختلفی به عواطف به‌عنوان عاملی اشاره کرده‌اند که در وجود یا عدم و میزان ظرفیت اخلاقی مصنوعات مدخلیت دارد. در این مقاله به رابطه میان عواطف و ظرفیت اخلاقی هوش مصنوعی خواهیم پرداخت. پرسش اصلی این است که عواطف در ارتقاء ظرفیت اخلاقی فناوری‌های هوش مصنوعی نقشی مثبت دارند یا منفی؟ چهار استدلال از جانب موافقین نقش مثبت عواطف در ظرفیت اخلاقی استخراج و صورت‌بندی می‌شود که شامل استدلال از طریق «حساسیت اخلاقی»، «عقلانیت محدود»، «برآورد خطر» و «مجازات‌پذیری» می‌شود. همچنین چهار استدلال از جانب مخالفین نقش مثبت عواطف در ظرفیت اخلاقی هوش مصنوعی ارائه می‌شود که شامل استدلال از طریق «ربایش عاطفی»، «سراب عواطف»، «پارادوکس انسان‌انگاری و انسانیت‌زدایی همزمان» و «مهارت‌زدایی اخلاقی» است. در نهایت، با شفاف‌سازی نقاط نزاع دو طرف بحث، دیدگاه‌های فوق مورد تحلیل قرار گرفته و چالش‌های اخلاقی پیاده‌سازی عواطف در هوش مصنوعی شرح داده می‌شود.

استناد: زرگر؛ زهرا. (۱۴۰۴). رابطه عواطف و ظرفیت اخلاقی در فناوری‌های هوش مصنوعی، *پژوهش‌های فلسفی*، ۱۹(۵۰)، ۱۹-۴۰.
https://doi.org/10.22034/jpiut.2024.63626.3875

ناشر: دانشگاه تبریز.

© نویسندگان.

مقدمه

هرچند این ایده که فناوری‌ها به‌لحاظ ارزشی خنثی هستند و اخلاقی یا غیراخلاقی بودن کاربرد فناوری صرفاً به شیوه استفاده کاربر از آن بازمی‌گردد، زمانی بین متخصصین و مردم طرفداران بسیاری داشت، اما امروزه در میان پژوهشگران فناوری کمتر کسی را می‌توان یافت که منکر ارزش‌باری اخلاقی فناوری‌ها باشد.[۱] درباره ارزش‌باری اخلاقی فناوری می‌توان پرسش‌های جالبی مطرح کرد؛ مانند این که بر اساس چه تعابیری از ارزش‌باری می‌توان ادعا کرد فناوری‌ها ارزش‌بار هستند، یا این که نسبت فناوری با ارزش‌های اخلاقی، عینی[۲]است یا برساختی[۳](یا چیزی مابین این دو). اگر از این مسائل بگذریم، با به‌رسمیت شمردن نقش فناوری در تحقق ارزش‌های اخلاقی، این پرسش مطرح می‌شود که چطور می‌توان از فناوری‌ها برای تحقق بیشتر اخلاق و ارتقاء سطح اخلاق بهره برد.

فیلسوفان متعددی درباره چگونگی تأثیر فناوری‌ها بر ارزش‌های اخلاقی سخن گفته‌اند. به‌طور کلی می‌توان این تأثیرگذاری را در سه حالت ارتباط انسان و فناوری در نظر گرفت: حالتی که در آن فناوری به مثابه بستر تجربه و عمل انسانی است؛ حالتی که در آن فناوری برای انسان به مثابه *دیگری* پدیدار می‌شود و جایگاه مخاطب عمل اخلاقی را دارد و حالتی که در آن فناوری خود یک عامل اخلاقی است.

حالت اول ارزش‌باری فناوری با آثاری چون مقاله «آیا مصنوعات سیاست دارند؟» لانگدون وینر مورد توجه قرار گرفت. او در این کتاب به پل‌های روگذر در لانگ آیلند نیویورک اشاره می‌کند که عمداً کم‌ارتفاع ساخته شده بودند تا دسترسی فقرا را که با اتوبوس‌های ارزان‌قیمت و مرتفع عبور و مرور می‌کردند به مراکز تفریحی مسدود کنند (وینر، ۱۹۸۰، ۱۲). پیتر پاول فربیک هم در کتاب *اخلاقی‌سازی فناوری* بر نقش طراحان در تعبیه ارزش‌های اخلاقی در فناوری تأکید کرده و می‌گوید: «طراحان به اخلاق مادّیت می‌بخشند» (فربیک، ۲۰۱۱، ۴۰)، اما از دید او نسبت فناوری با عمل اخلاقی نه نسبتی تعیّن‌بخش، بلکه از جنس وساطت است. فناوری‌ها در اعمال و تصمیمات انسانی وساطت می‌کنند و از این جهت جنبه‌ای اخلاقی دارند. برای مثال فناوری سونوگرافی هنگامی که برای مشاهده جنین به کار می‌رود، از یک سو جنین را به‌عنوان بیمار بازنمایی کرده و والدین را ترغیب می‌کند در صورت رؤیت نقص جدی به فکر ختم بارداری باشند، از سویی دیگر با محسوس کردن جنین و ایجاد پیوند عاطفی بین او و والدین آنها را از سقط جنین بازمی‌دارد. در اینجا این فناوری والدین را به سمت یک عمل اخلاقی یا غیراخلاقی سوق نمی‌دهد، بلکه شرایط تصمیم‌گیری آنها را شکل داده و متحول می‌کند (فربیک، ۲۰۱۱، ۲۶).

در حالت دوم، فناوری به شبکه ارتباطات اجتماعی وارد می‌شود و برای فرد جایگاهی مشابه جایگاه یک انسان دیگر پیدا می‌کند. این حالت با ظهور فناوری‌های جدید مانند فضای مجازی و ربات‌های اجتماعی مورد توجه قرار گرفته‌است. برخی فیلسوفان معتقدند در صورتی که رابطه با یک مصنوع، حالات و عواطفی مشابه رابطه با یک انسان را در فرد برانگیزد، نحوه عمل کردن در این رابطه می‌تواند مشمول ملاحظات اخلاقی شود و مصنوعات در این رابطه نوعی شأنیت اخلاقی دارند. برای مثال کاکلبرگ در تحلیل بازی‌های رایانه‌ای خشن به این نکته اشاره می‌کند که خشونت‌ورزی فعّال نسبت به آوارتارها در فضایی که بسیار شبیه به جهان واقعی است بر

---

[۱] هرچند جوزف پیت یکی از این افراد است (پیت، ۲۰۱۴).

[۲] Objective

[۳] Constructed

شخصیت اخلاقی فرد اثر می‌گذارد و فضیلت همدلی را در او تضعیف می‌کند (کاکلبرگ، ۲۰۰۷، ۲۴-۲۵). در نمونه‌ای دیگر اسپارو با اشاره به ربات‌های اجتماعی که به‌عنوان جایگزین حیوان خانگی استفاده می‌شوند، بدرفتاری با این ربات‌ها -لگد زدن به یک سگ ربات- را غیراخلاقی دانسته و آن را باعث ایجاد رذیلت اخلاقی در فرد می‌داند (اسپارو، ۲۰۱۶).

حالت سوم، حالتی است که در آن مصنوع به منزله عامل اخلاقی دیده می‌شود. در این حالت مصنوع به‌عنوان مصدر افعالی در نظر گرفته می‌شود که به‌لحاظ اخلاقی قابل ارزش‌گذاری هستند. این حالت از ارزش‌باری فناوری‌ها در پی توسعه فناوری‌های هوش مصنوعی یادگیرنده که درجه‌ای از خودمختاری را از خود بروز می‌دهند موضوعیت پیدا کرده‌است. با این حال، در بین فیلسوفان درباره امکان‌پذیری عاملیت اخلاقی مصنوعی[۱] اختلاف نظرهایی جدی وجود دارد. این در حالی است که پاسخ پرسش‌های مهمی درباره طراحی فناوری‌های هوش مصنوعی، تصمیم‌گیری برای واگذاری وظایف مختلف به آنها و توزیع مسئولیت اعمال فناورانه به پاسخ ما به این پرسش بستگی دارد.

از بین حالات فوق، عاملیت اخلاقی حالتی اختصاصی‌تر برای فناوری‌های هوش مصنوعی است، چرا که بستر تجربه و عمل بودن و مخاطب عمل اخلاقی بودن، تا حدودی برای سایر مصنوعات نیز قابل طرح است. بر این اساس، در این مقاله با تمرکز بر مسئله عاملیت اخلاقی هوش مصنوعی، بر رابطه عواطف و ظرفیت اخلاقی مصنوعات متمرکز می‌شویم. عواطف اغلب به‌عنوان مؤلفه‌ای غیرمستقیم و ضمنی در شرایط لازم و کافی عاملیت اخلاقی محسوب شده، یا جزو عوامل دخیل در عاملیت اخلاقی به‌حساب آمده‌اند. با این حال، درباره نقش مثبت یا منفی عواطف در توسعه ظرفیت‌های اخلاقی هوش مصنوعی دیدگاه‌های متفاوتی وجود دارد. پرسش اصلی که به دنبال پاسخ آن هستیم این است که آیا عواطف در ارتقاء ظرفیت‌های اخلاقی مصنوعات نقشی مثبت دارند یا منفی؟ برای پاسخ به این پرسش، استدلال‌هایی به نفع و علیه نقش مثبت عواطف در ارتقاء ظرفیت اخلاقی استخراج، ارائه و تحلیل می‌شود.

در قسمت بعدی مقاله، چشم‌اندازی از مباحث مربوط به عاملیت اخلاقی ترسیم می‌شود که شامل ارائه تعاریف مختلف عاملیت و مرور دیدگاه‌های اصلی درباره آن است. در قسمت سوم، زوایای مختلف رابطه عواطف و ظرفیت اخلاقی مصنوعات مورد بررسی قرار می‌گیرد. در قسمت چهارم، استدلال‌های مدافع نقش مثبت عواطف در ظرفیت‌های اخلاقی مصنوعات و در قسمت پنجم استدلال‌های مدافع نقش منفی عواطف در ظرفیت اخلاقی مصنوعات شرح داده‌می‌شوند. نهایتاً در قسمت ششم، نقاط نزاع و چالش‌های پیش روی طراحی مصنوعات عاطفه‌مند مورد تحلیل قرار می‌گیرد.

۱. عاملیت اخلاقی مصنوعی

عاملیت در تعبیری عام به معنای توانمندی یک هویت بر انجام عمل است. به‌طور معمول میان «رفتار»[۲] (کنش‌هایی که صرفاً بر اساس روابط علّت و معلولی علمی توضیح داده می‌شوند) و «عمل»[۳] (کنش‌هایی که علاوه بر روابط علت و معلولی، با اشاره به قصد کنش‌گر توضیح داده می‌شوند) تفاوت وجود دارد. «عاملیت» مربوط به توانمندی یک هویت بر انجام «عمل» است. هنگامی که عمل مورد نظر ما دارای جنبه‌ای اخلاقی باشد، صحبت از عاملیت اخلاقی به میان می‌آید.

---

[۱] Artificial Moral Agency

[۲] Behavior

[۳] Action

پرسش از عاملیت اخلاقی مصنوعات، در سه بعد وجودشناختی، معرفت‌شناختی و عملی امتداد پیدا می‌کند[۱]؛ نخست این که چه ویژگی‌هایی قوام‌بخش شروط لازم و کافی برای داشتنِ عاملیت اخلاقی هستند؟ دوم این که چطور می‌توانیم نسبت به وجود یا عدم این ویژگی‌ها در یک هویت معرفت پیدا کنیم؟ و سوم این که با فرض امکان‌پذیری عاملیت اخلاقی مصنوعات، چطور می‌توان ظرفیت‌های اخلاقی را در مصنوعات ایجاد کرد؟

برخی فیلسوفان پرداختن به مسئله عاملیت را از پرسش اول آغاز کرده‌اند. برای مثال فلوریدی و ساندرس سه ویژگی را به‌عنوان شروطی معرفی می‌کنند که در صورت برآورده شدن در یک هویت، آن را دارای جایگاه عاملیت اخلاقی می‌کنند: تعاملی بودن، خودمختاری و انطباق‌پذیری.[۲] تعاملی بودن به این معناست که عامل و محیط پیرامونش بتوانند با یکدیگر کنش داشته‌باشند. خودمختاری به معنای این است که عامل بتواند تغییر حالت دهد بی‌آن که این تغییر، پاسخی مستقیم به تعامل باشد. پس طبق این شرط یک عامل خودمختار باید اقلاً دو حالت ممکن داشته‌باشد و دارای سطح خاصی از پیچیدگی و استقلال از پیرامون خود باشد. منظور از انطباق‌پذیری نیز آن است که تعاملات عامل بتوانند قوانینِ گذاری را که بر اساس آن حالاتش را تغییر می‌دهد، عوض کنند. این ویژگی تضمین می‌کند که عامل حالت عملکردش را به‌طریقی که بسیار وابسته به تجربیاتش است بیاموزد (فلوریدی و ساندرس، ۲۰۰۴، ۳۵۸–۳۵۷).

سالینز خودمختاری، قصدمندی و مسئولیت‌پذیری را شروط سه‌گانه عاملیت معرفی می‌کند. تعبیر مدنظر سالینز از خودمختاری، تعبیری مهندسی است؛ یعنی استقلال از برنامه‌نویس، اپراتور و کاربر. منظور از قصدمندی نیز این است که اگر تعامل پیچیده برنامه‌های ربات و محیط باعث شود ماشین به‌طریقی عمل کند که به‌لحاظ اخلاقی زیان‌بار یا سودمند باشد و این اعمال *ظاهراً* محصول رایزنی و محاسبات باشند، می‌توان گفت عمل ماشین قصدمندانه است. مطابق شرط مسئولیت‌پذیری، وقتی یک ربات نقش‌هایی اجتماعی را بر عهده داشته‌باشد که دربردارنده مسئولیت‌هایی اخلاقی باشند و تنها راه معنابخشی به عملکردش این باشد که فرض کنیم ربات «باور» دارد نسبت به انجام وظایفش مسئول است، می‌توانیم بگوییم به‌اندازه کافی مسئولیت‌پذیر است. بر اساس این سه شرط، یک مصنوع برای این که عامل اخلاقی به حساب بیاید نیازی به این ندارد که دارای ذهن یا آگاهی باشد (سالینز، ۲۰۰۶، ۲۸–۲۹).

هرچند فلوریدی، ساندرس و سالینز از امکان عاملیت اخلاقی مصنوعات دفاع می‌کنند، اما برخی فیلسوفان با ایشان هم‌نظر نیستند. هیما نماینده این گروه است. او ظرفیت انتخاب آزادانه اعمال خود و ظرفیت تشخیص تفاوت میان درست و غلط را به‌عنوان شروط لازم و مجموعاً کافی برای عاملیت اخلاقی معرفی می‌کند. منظور از انتخاب آزادانه، انتخابی است که مبتنی بر تأمّل بوده و خروجی آن توسط چیزی بیرون از عامل متعیّن نشود. در مورد شرط معرفت اخلاقی نیز هیما منظور خود را از معرفت چیزی ضعیف‌تر از باور صادق موجه در نظر گرفته و آن را صرفاً وجود نوعی روش‌شناسیِ عمدتاً اعتمادپذیر برای بازشناسی درست از غلط می‌داند. به‌طور کلی از نظر هیما شرایط عاملیت اخلاقی، شرایطی است که یک عامل را برای درگیر شدن در استدلال اخلاقی و عمل بر اساس آن، توانمند می‌کند، اما برای درگیر شدن در استدلال اخلاقی، فهمی حداقلی از مفاهیم اخلاقی همچون خوب، بد، الزام‌آور، غلط و مجاز، لازم است. همچنین، داشتنِ دریافتی از اصول اخلاقی پایه‌ای – مثل اصل پرهیز از آسیب– ضرورت دارد (هیما، ۲۰۰۹، ۱۳–۱۶). بر

---

[۱] البته برخی فیلسوفان در ارتباط با وضعیت اخلاقی مصنوعات، رویکردی رابطه‌ای را برگزیده‌اند که طبق آن نه ویژگی‌های وجودی یک هویت، بلکه ارتباطی که میان فرد و او شکل می‌گیرد در وضعیت اخلاقی هویت برای فرد دخیل است (کاکلبرگ، ۲۰۱۴؛ کاپوچو و همکاران، ۲۰۱۹).

[۲] Interactivity, Autonomy, and Adaptability.

این اساس، هیما «آگاهی» را به‌عنوان یک شرط ضمنی و ضروری برای عاملیت اخلاقی معرفی می‌کند. چرا که اولاً داشتنِ قصد ملازمِ داشتنِ آگاهی است؛ ثانیاً فرآیند استدلال فرآیندی است که در آن دلایل فهمیده می‌شوند و چنین فرآیندی لاجرم آگاهانه است (هیما، ۲۰۰۹، ۱۶-۱۷).

گروهی دیگر از فیلسوفان با همدلی نسبی با دلایل هر دو گروه در تأیید یا ردّ عاملیت اخلاقی مصنوعات، راه دیگری در پیش گرفته‌اند و به جای آن که درباره مفهومی یکتا از عاملیت اخلاقی بحث کنند، توصیفی ریزبافت‌تر از عاملیت اخلاقی ارائه داده که شامل چندین سطح از عاملیت می‌شود و برای هر یک از این سطوح شروطی را برشمرده‌اند. برای مثال مور چهار سطح از عاملیت اخلاقی را از هم متمایز می‌کند: عامل اخلاقی اثرگذار[۱] عامل اخلاقی ضمنی[۲] عامل اخلاقی آشکار[۳]و عامل اخلاقی کامل.[۴] هر ماشینی که بر اساس پیامدهای اخلاقی عملکردش مورد ارزیابی قرار بگیرد، دارای عاملیت اخلاقی اثرگذار است. در سطح بالاتر، ماشین‌هایی هستند که به گونه‌ای طراحی شده‌اند که فاقد آثار اخلاقی منفی باشند، یعنی ایمنی و اعتمادپذیری در فرآیند طراحی آنها لحاظ شده‌است. این ماشین‌ها عاملیت اخلاقی ضمنی دارند. سطح بعدی عاملیت، به ماشین‌هایی اختصاص دارد که برنامه‌ریزی آنها به‌نحوی است که با استفاده از مقولات اخلاقی می‌توانند استدلال کنند و دارای عاملیت اخلاقی آشکار هستند. در بالاترین سطح عاملیت اخلاقی، عاملان کامل هستند؛ یعنی عاملانی که داوری‌های اخلاقی صریحی انجام می‌دهند و صلاحیت توجیه تصمیمات اخلاقی‌شان را دارند. این سطح از عاملیت مستلزم ظرفیت آگاهی، قصدمندی و اراده آزاد است. ماشین‌ها، دو مرتبه اول عاملیت اخلاقی را به‌سادگی می‌توانند احراز کنند، اما احراز مرحله سوم با چالش‌های عملی بسیاری همراه است. درحالی‌که نیل به سطح چهارم و عامل اخلاقی کامل بودن، برای هیچ ماشینی میسّر نیست (مور، ۲۰۰۶، ۱۹-۲۰).

در رویکردی دیگر نسبت به عاملیت اخلاقی مصنوعات، پیشنهاد می‌شود معضل نظری تعریف شرایط عاملیت به‌طریقی دور زده شود و به جای تمرکز بر این مسئله، رویکردی هنجاری درباره عاملیت مصنوعات در پیش گرفته شود. در چنین رویکردی فرض بر این است که عاملیتی شراکتی بین مصنوعات و انسان‌ها وجود دارد و این شراکت شکل‌پذیر است. بنابراین، به جای پرسش از امکان‌پذیری عاملیت مصنوعات، بر بایدها و نبایدهای عاملیت شراکتی انسان–مصنوع تمرکز می‌شود. در این رویکرد ملاحظات اخلاقی در ترکیب‌بندی شراکت میان انسان و مصنوع، واگذاری وظایف به مصنوعات دارای هوش مصنوعی بر اساس توانمندی‌ها و ایجاد قابلیت‌های اخلاقی در آنها مد نظر قرار خواهدگرفت (بهدادی و مونثه، ۲۰۲۰، ۲۱۶-۲۱۷).

در مجموع، اختلاف نظر اساسی فیلسوفان درباره عاملیت اخلاقی مصنوعات مربوط به افق نهایی است که برای قابلیت مصنوعات در تولید اعمال اخلاقی می‌توان تصور کرد.[۵] با این حال، این که فناوری‌های هوش مصنوعی در تولید اعمال اخلاقی تفاوت چشم‌گیری با سایر مصنوعات دارند، مورد تأیید اغلب فیلسوفان است. این امر، مسئله چگونگی توسعه ظرفیت‌های اخلاقی[۶] در هوش مصنوعی –با تمام ابعاد نظری و عملیاتی‌اش– به پیش می‌کشد. یکی از موضوعات مرتبط به ظرفیت‌های اخلاقی هوش مصنوعی، عواطف

---

[۱] Ethical Impact Agent

[۲] Implicit Ethical Agent

[۳] Explicit Ethical Agent

[۴] Full Ethical Agent

[۵] برای مرور دیدگاه‌هایی بیشتر درباره عاملیت اخلاقی مصنوعات این آثار را ببینید: (جانسون و نورمن، ۲۰۱۴)، (برِی، ۲۰۱۴)، (پاورز، ۲۰۱۳)، (فرنکلین و گراسر، ۱۹۹۶)، (نیلهوم، ۲۰۱۹)، (برینگزیورد، ۲۰۰۷).

[۶] Moral Capacities

است. اهمیت عواطف در ارتباط با ظرفیت اخلاقی مورد تأکید محققان بسیاری قرار گرفته است، اما چیستی رابطه عواطف با ظرفیت اخلاقی موضوعی ذوابعاد و محل مناقشات گوناگون بوده‌است.

۲. زوایای گوناگون نسبت عواطف و ظرفیت اخلاقی مصنوعات

بحث درباره نسبت عواطف و ظرفیت اخلاقی مصنوعات، از یک سو به امکان پیاده‌سازی عواطف یا شبه‌عواطف در مصنوعات (هوش مصنوعی) بازمی‌گردد. برخی فیلسوفان اساساً چنین امکانی را رد می‌کنند و برخی دیگر با لحاظ کردن راهبردهای عملیاتی مختلف (مانندOCC[۱]) آنها از نظر موفقیت در پیاده‌سازی عواطف ارزیابی می‌کنند. جهت‌گیری محققان نسبت به این مسئله، تا حد زیادی بستگی به این دارد که درباره عواطف چه نظریه‌ای را برگزیده‌اند.

پرینز نظریه‌های ارائه‌شده درباره عواطف را به چند دسته تقسیم کرده و حالات ترکیبی آنها را نیز برمی‌شمرد. طبق نظریه احساسات[۲] عواطف آن چیزی هستند که از تغییرات بدنی حس می‌کنیم. نظریه‌پرداز پیشتاز این دیدگاه، جیمز[۳]است. جیمز عواطف را احساس تغییراتی بدنی چون تپش قلب و تعریق می‌داند و به‌لحاظ زمانی تغییرات بدنی را مقدم بر عواطف می‌داند. در این رویکرد حس کردن تغییر بدنی، مقوّم یک انگیزش عاطفی است (پرینز، ۲۰۰۴، ۵).

رویکرد داماسیو[۴]شباهت‌های مهمی با رویکرد جیمز دارد. داماسیو نیز منشاء عواطف را تغییرات بدنی می‌داند با این تفاوت که تغییرات عصبی و شیمیایی مغزی را نیز در زمره این تغییرات و بعضاً مقدم بر آنها می‌داند. با این حال، او معتقد است در بسیاری از موارد تغییرات حالات مغزی، بدن را دور می‌زند و به علت تغییرات شیمیایی عصبی در مغز عواطف ایجاد می‌شوند، بدون این که تغییرات بدنی خاصی رخ دهد. همچنین، ممکن است گاهی تغییرات مغزی ملازم با یک انگیزش عاطفی رخ دهند، اما عاطفه وارد سطح آگاهی فرد نشود. بنابراین در رویکرد داماسیو تغییرات بدنی مقوم اصلی انگیزش عاطفی است و حس شدن آن شرط لازم نیست. به همین دلیل به این رویکرد بدنی[۵]گفته می‌شود (پرینز، ۲۰۰۴، ۶).

رویکردی دیگر، نظریه رفتاری است که گیلبرت رایل[۶]آن را نمایندگی می‌کند. از دید رایل عواطف، گرایش‌هایی برای عمل کردن به‌طریقی خاص هستند. مثلاً وحشت‌زده بودن یعنی گرایش داشتن به بروز رفتارهایی مثل منقبض کردن خود، فریاد کشیدن، یا به یاد نیاوردن راه فرار (پرینز، ۲۰۰۴، ۷).

برخی محققان به جنبه‌های شناختی عواطف توجه کرده‌اند. ایشان عواطف را با پیمانه‌های شناختی ذهن مثل حافظه، توجه و استدلال، در تعاملی نظام‌مند می‌بینند. مثلاً عواطف موجب می‌شوند به یاد آوردن رویدادهایی که در آنها حالتی عاطفی داشته‌ایم که اکنون داریم، برای ما ساده‌تر شود. عواطف مثبت به ما کمک می‌کنند خلاقانه، با ذهن گشوده و خوش‌بینانه دلیل‌آوری کنیم و عواطف منفی محدوده نگرشمان را تنگ‌تر کرده و توجه ما را به تهدیدات جلب می‌کنند. پرینز این نظریه را نظریه حالت پردازشگر نامیده و

---

[۱] OCC یک مدل شناختی محاسباتی درباره عواطف است که توسط Orteny، Clore و Colins ابداع و بر اساس حروف اول نام ایشان نامگذاری شده‌است (اُرتنی و همکاران، ۲۰۲۲).

[۲] Feeling Theory of Emotion

[۳] William James

[۴] Antonio Damasio

[۵] Somatic Theory of Emotion

[۶] Gilbert Ryle

اوتلی و جانسون لیرد[۱]را به‌عنوان مبدعانش معرفی می‌کند (پرینز، ۲۰۰۴، ۸). بعضی فیلسوفان عواطف را صرفاً شناختی می‌دانند. از دید ایشان عواطف با افکار این‌همان هستند. مثلاً ترس، باور به وجود خطر و میل به اجتناب از خطر است. بدفورد[۲]و سولومون[۳]از رویکردی شناختی درباره عواطف دفاع می‌کنند. پرینز این رویکرد را شناختی محض[۴]می‌نامد (پرینز، ۲۰۰۴، ۸).

این پنج نظریه، در ارتباط با امکان عاطفه‌مندی هوش مصنوعی دلالت‌های متفاوتی دارند. والاچ و الن در تحلیل این رابطه به این نکته اشاره می‌کنند که نظریه‌های احساسات و رفتاری، عواطف را خروجی می‌دانند درحالی‌که نظریه‌های حالت پردازشگر و شناختی، عواطف را از جنس فرآیند می‌دانند. به همین دلیل نظریه‌های حالت پردازشگر و شناختی چشم‌انداز بهتری برای پیاده‌سازی توسط هوش مصنوعی ترسیم می‌کنند. همچنین نظریه بدنی، به دلیل این که مشخص نیست تا چه حد ویژگی‌های فیزیولوژیک قابل پیاده‌سازی توسط هوش مصنوعی باشد و جنبه‌های منحصراً زیستی تا چه حد در ایجاد عواطف حیاتی باشند، نویدبخش پیاده‌سازی عواطف در هوش مصنوعی نیست (والاچ و الن، ۲۰۰۹، ۱۴۴-۱۴۵).

جنبه بحث‌برانگیز دیگر درباره نسبت عواطف و ظرفیت اخلاقی، میزان اهمیت عواطف در نسبت با ظرفیت‌های اخلاقی است. برخی فیلسوفان عواطف را رکنی اساسی برای وجود ظرفیت اخلاقی می‌دانند درحالی‌که برای برخی دیگر عواطف مدخلیتی واسطه‌ای در ظرفیت اخلاقی دارد. موضع فیلسوفان در قبال این مسئله، تا حدی به نظریه مختارشان درباره اخلاق باز می‌گردد. برای مثال در نسخه‌هایی از اخلاق فایده‌گرایانه که لذت و رنج ایجاد شده را ملاک ارزیابی یک عمل می‌داند، درجاتی از عاطفه‌مندی و قابلیت همدلی لازمه سطوح بالای ظرفیت اخلاقی است. از منظر اخلاق فضیلت‌گرا، عاطفه‌مندی به‌نحوی غیرمستقیم‌تر و در فرآیند کسب فضایل یا رذایل می‌تواند مرتبط محسوب شود، اما از منظر اخلاق رواقی، یا اخلاق کانتی (وظیفه‌گرایانه) عواطف نقش مثبتی در عملکرد اخلاقی ندارند؛ یعنی یا نامرتبط شمرده می‌شوند یا نقشی منفی دارند.

یکی دیگر از جنبه‌های جالب رابطه عواطف و ظرفیت اخلاقی، پرسش از مثبت یا منفی بودن نقش عواطف در ارتقاء ظرفیت اخلاقی مصنوعات است. برخی فیلسوفان برای عواطف نقشی سازنده در ظرفیت اخلاقی قائلند تا حدی که وجود عواطف را شرط ضروری برای درک، تصمیم‌گیری و عمل اخلاقی می‌دانند، اما برخی دیگر بر تأثیرات منفی عواطف روی عاملیت اخلاقی تأکید کرده و به همین سبب معتقدند می‌توان در چشم‌انداز توسعه مصنوعات فاقد عواطف، به وجود عاملانی اخلاقی حتی به‌مراتب شایسته‌تر از انسان‌ها امید داشت. موضع فیلسوفان در قبال این موضوع، به دیدگاه‌های ایشان درباره مسائلی دیگر چون نظریه عواطف، تعریف عاملیت و شأنیت اخلاقی، شرایط احراز عاملیت اخلاقی و قابلیت روش‌های مهندسی برای پیاده‌سازی عواطف بستگی دارد. در دو قسمت بعدی، دیدگاه‌های اصلی درباره نقش مثبت یا منفی عواطف در ایجاد و توسعه ظرفیت‌های اخلاقی مورد بررسی قرار می‌گیرد.

۳. عواطف به مثابه عاملی مثبت در ارتقاء ظرفیت‌های اخلاقی

در این بخش به چهار دیدگاه اصلی در دفاع از نقش مثبت عواطف در ارتقاء ظرفیت‌های اخلاقی پرداخته می‌شود که شامل استدلال‌های «عواطف و حساسیت اخلاقی»، «عواطف و عقلانیت محدود»، «عواطف و برآورد خطر» و «عواطف و مجازات‌پذیری» هستند. هرچند

---

[۱] Oatley and Johnson Laird
[۲] Bedford
[۳] Solomon
[۴] Pure Cognitive

محققینی که در این قسمت آراءشان مورد بررسی قرار می‌گیرد نقش مثبت عواطف را در ظرفیت اخلاقی تأیید می‌کنند، اما درباره امکان‌پذیری عاملیت اخلاقی هوش مصنوعی (با توجه به نسبت عواطف و ظرفیت اخلاقی) هم‌نظر نیستند.

۱-۳. عواطف و حساسیت اخلاقی

در میان موقعیت‌های مختلفی که در آن عمل می‌کنیم، برخی موقعیت‌ها اخلاقی و برخی دیگر نااخلاقی هستند. برای مثال در شرایط عادی، انتخاب رنگ سبز یا آبی برای لباس موقعیتی نااخلاقی است، اما انتخاب استفاده از کیسه پلاستیکی یا پارچه‌ای برای خرید، موقعیتی اخلاقی است. هر عامل اخلاقی برای این که عاملی شایسته باشد باید بتواند اهمیت اخلاقی یک موقعیت را درک کند، به این ویژگی حساسیت اخلاقی[۱] گفته می‌شود. بسیاری از محققین وجود حساسیت اخلاقی را مستلزم نوعی عاطفه‌مندی می‌دانند؛ گویی این عواطف هستند که با برجسته ساختن برخی موقعیت‌ها و پدیدار کردن آنها به‌نحوی خاص، ما را متوجه اهمیت اخلاقی آن موقعیت می‌کنند.

بر این اساس، برخی محققین معتقدند از آنجا که هوش مصنوعی نمی‌تواند به‌لحاظ عاطفی درگیر شود، نباید در وظایفی مستلزم حساسیت اخلاقی بالا است، هوش مصنوعی را جایگزین انسان کرد. فاریسکو و همکارانش می‌گویند مغز انسان عاطفی است به این معنا که ترجیحاتی دارد که عملکرد او را معین می‌کند و به‌نحوی مهم با پاسخ‌هایی عاطفی مثل حساسیت نسبت به پیام‌های پاداش سامان یافته‌است، اما هوش مصنوعی به این معنا عاطفی نیست. از نظر فاریسکو و همکارانش عواطف می‌توانند در ارزیابی کیفیت اخلاقی اعمال نقشی مثبت داشته‌باشند و حتی عملی که با درگیری عاطفی مناسب انگیزه‌دهی و همراه شده‌باشد نسبت به عملی که فاقد چنین درگیری عاطفی است ارزش اخلاقی بالاتری دارد. این موضوع خصوصاً در ارتباط با برخی اَعمال بهتر قابل رؤیت است. مثلا پزشکی که وظیفه طبابت خود را همراه با همدلی و درگیری عاطفی انجام می‌دهد به‌لحاظ اخلاقی از پزشکی که بدون درگیری عاطفی به درمان مریض رسیدگی می‌کند برتر است. علت این برتری این است که درگیری عاطفی، توجه ما را به سمت موضوعاتی جلب می‌کند که باید به آن توجه کرد. همدلی همچون یک قطب‌نمای اخلاقی عمل می‌کند و ما را نسبت به نیازهای دیگران بینا می‌کند. به این ترتیب، فقدان درگیری عاطفی هوش مصنوعی باعث می‌شود فاقد برخی ظرفیت‌های اخلاقی باشد و برای انجام برخی اعمال مناسب نباشد (فاریسکو و همکاران، ۲۰۲۰، ۲۴۱۹).

ولیز نیز از محققان هوش مصنوعی است که ادعا می‌کند فقدان عواطف، باعث می‌شود فناوری‌های هوش مصنوعی عاملیت اخلاقی نداشته‌باشند. ولیز عواطف را دارای نقشی مهم در حساسیت اخلاقی توصیف می‌کند. او می‌گوید حس‌مندی همچون یک آزمایشگاه اخلاقی درونی است که اعمال ما را هدایت می‌کند. داشتن تجربه آسیب‌دیدگی، رنج و درد و لذت و رضایت، به ما کمک می‌کند از این که اعمال ما چه تأثیری روی دیگران می‌گذارند تصوری پیدا کنیم. هرچه تجربیات ما به دیگری نزدیک‌تر باشد، شکاف همدلی میان ما و او کمتر می‌شود. از آنجا که الگوریتم‌ها فاقد چنان تجاربی هستند، نمی‌توان آن‌ها را دارای شروط اولیه عاملیت اخلاقی دانست (ولیز، ۲۰۲۱، ۴۹۳).

گویندا‌راجولو و همکارانش با رویکردی فضیلت‌گرایانه مسئله حساسیت اخلاقی را در ارتباط با الگوهای ممتاز اخلاقی در نظر می‌گیرند. در برخی از قرائت‌های اخلاق فضیلت‌گرا، الگوهای اخلاقی در جهت‌دهی فرد نسبت به اعمال اخلاقی نقشی کلیدی دارند:

---

[۱] Moral Sensitivity

در هر موقعیت، عملی اخلاقی است که اگر یک الگوی اخلاقی در آن موقعیت می‌بود، آن را انجام می‌داد. پس برای یک عامل اخلاقی مهم است که بتواند الگوی اخلاقی را شناسایی کند. گویندراجولو و همکارانش معتقدند شناسایی الگوی اخلاقی، از یک انگیزش عاطفی شروع می‌شود: تحسین. پس اگر بنا باشد یک فناوری هوش مصنوعی بر اساس اخلاق فضیلت‌گرا یک عامل اخلاقی محسوب شود، باید بتواند انگیزش تحسین را تجربه کرده و بر اساس آن الگوهای اخلاقی را شناسایی کند. آنها بر خلاف ولیز، نسبت به پیاده‌سازی عواطف در هوش مصنوعی خوش‌بین هستند و در پروژه خود صوری‌سازی و پیاده‌سازی انگیزش عاطفی تحسین در هوش مصنوعی را دنبال می‌کنند (گویندراجولو و همکاران، ۲۰۱۹، ۲ و ۵).

۲-۳. عواطف و عقلانیت محدود

یک عامل هوش مصنوعی را در نظر بگیرید که در یک موقعیت اخلاقی قرار گرفته و بناست مطابق اخلاق فایده‌گرایانه عمل کند. برای مثال خودرویی خود-ران که در حال حرکت است و ناگهان گروهی از کودکان وسط خیابان می‌پرند و در مسیر خودروها قرار می‌گیرند. پردازشگر خودرو در چنین موقعیتی باید بتواند راه‌های مختلف ممکن برای عمل را در نظر بگیرد، فواید و مضرّات هر راه را محاسبه کند و در نهایت با به کار بستن قاعده اصلی فایده‌گرایی -یعنی بیشینه کردن فایده- یک تصمیم اخلاقی بگیرد. در این فرآیند عامل مصنوعی با یک چالش مهم مواجه است: چه اطلاعاتی به تصمیم‌گیری اخلاقی او مرتبط هستند و تا کجا باید جستجوی اطلاعات و پردازش آنها را ادامه دهد؟ درباره این موقعیت دامنه نامحدودی از اطلاعات را می‌توان جمع‌آوری کرد؛ مانند ساعت وقوع رویداد، دما و رطوبت هوا در محل رویداد، سن و جنسیت کودکان، سرعت حرکت آنها، مدل ماشین‌های حاضر در خیابان، وزن سرنشینان خودرو و ... . کدام‌یک از این اطلاعات به محاسبه فایده و مضرات راه‌های مختلف مرتبطند؟ دامنه محاسبه پیامدهای گزینه‌های مختلف (توقف ناگهانی، تغییر مسیر به سمتی که با کودکان کمتری برخورد کند، یا تغییر مسیر به سمت خودروهای دیگر و ....) را تا کجا باید ادامه داد؟ اگر یک عامل در چنین موقعیتی نتواند اطلاعات مربوط و دامنه محاسباتی را محدود کند، نمی‌تواند در زمان مناسب یک تصمیم اخلاقی بگیرد.

در انسان‌ها این محدودسازی، توسط عواطف رخ می‌دهد. سایمون از اصطلاح عقلانیت محدود[۱] برای توصیف این ویژگی تفکر انسان‌ها استفاده می‌کند. انسان‌ها معمولاً در زمان اندک و با اتکا به اطلاعات محدودی تصمیم می‌گیرند. عقلانیت محدود به آنها اجازه می‌دهد در موقعیت تصمیم‌گیری بتوانند به سرعت و با قوانینی سرانگشتی، بهترین راه را پیدا کنند. سایمون از کلمه *رضایت‌بخش-کافی*[۲] استفاده می‌کند (سایمون، ۱۹۶۷، ۳۲). تا این ایده را مطرح کند که چنین تصمیماتی ممکن است تصمیمی نباشند که از دید یک عامل عقلانی نامحدود بهینه به نظر می‌رسد، اما بر اساس سنجه خود تصمیم‌گیرنده، به‌قدر کافی خوب هست (والاچ و الن، ۲۰۰۹، ۱۴۷).

داماسیو بر اساس دیدگاه جسمی خود راجع به عواطف انسانی، میان دو دسته عواطف تمایز قائل می‌شود: دسته اول واکنش‌های بدنمند و غریزی و دسته دوم، سازوکارهایی شناختی که می‌توانند از عواطف دسته اول برای هدایت رفتار از طریق تفکر و تأمّل استفاده کنند. داماسیو این دسته را عواطف ثانویه می‌نامد. عواطف اولیه به موجود زنده کمک می‌کند فرآیندهای تصمیم‌گیری کُند را که ممکن است بقائش را به خطر بیندازند دور بزند، اما عواطف ثانویه برای تصمیم‌گیری در مورد اهداف پیچیده‌تر مناسب هستند. داماسیو معتقد

---

[۱] Bounded Rationality

[۲] Satisficing

است در تصمیم‌های پیچیده که در آن نمی‌توان از نتایج مطمئن بود یا تفاوت میان گزینه‌ها واضح نیست، شاخص جسمی به‌ازای برآوردی از کلیه پیامدهای مثبت و منفی ممکن برای یک عمل، یک احساس کلی تولید می‌کند. این احساس به انتخاب ما جهت می‌دهد و عواطفی که ایجاد می‌شود گزینه‌های مختلف را ارزش‌گذاری می‌کنند. به این ترتیب عواطف در انتخاب مسیر عمل از بین گزینه‌های گوناگون نقش‌آفرینی می‌کنند (والاچ و الن، ۲۰۰۹، ۱۴۷–۱۴۹).

۳–۳. عواطف و برآورد خطر

تشخیص و برآورد میزان خطر، قابلیتی است که می‌تواند در خدمت ارتقاء ظرفیت اخلاقی باشد. این قابلیت باعث می‌شود عامل بتواند اعمال خود را به شکلی جهت‌دهی کند که به کمترین خطر احتمالی منتهی شود. سابینه روزر در آثار متعددی به نقش عواطف در برآورد خطر می‌پردازد (روزر، ۲۰۰۹؛ روزر، ۲۰۱۰؛ روزر، ۲۰۱۲). هرچند مسئله هوش مصنوعی و عاملیت اخلاقی مصنوعات مورد توجه روزر نبوده‌است، اما از استدلال‌های او درباره نقش شناختی عواطف در تعیین خطر می‌توان در بحث درباره ظرفیت‌های اخلاقی مصنوعات نیز استفاده کرد. به‌طور خاص، روزر درباره چگونگی ایفای نقش عواطف در طراحی مصنوعات تأمّل می‌کند و با توجه به این که برخی در چشم‌انداز توسعه هوش مصنوعی، تولید و طراحی مصنوعات دیگر به دست هوش مصنوعی را پیش بینی می‌کنند، بحث روزر می‌تواند در ارتباط با هوش مصنوعی نیز دلالت‌های جالبی داشته‌باشد.

روزر بحث خود را از اینجا آغاز می‌کند که طبق تحقیقات تجربی انجام شده، برای مردم معمولی، عواطف نقشی مهم در برآورد خطر یک موقعیت دارند، اما برای بسیاری از اندیشمندان، این نقش‌آفرینی عقلانی نیست. اغلب محققینی که روی رابطه خطرپذیری و عواطف کار می‌کنند نظریه فرآیندهای دوگانه را مفروض می‌گیرند که طبق آن ما با دو سامانه متمایز واقعیت را می‌فهمیم: سامانه اول که ناخودآگاه، سریع، شهودی و عاطفی است و سامانه دوم که آگاهانه، کُند، تحلیلی و عقلانی است. روزر این تقسیم‌بندی را نابسنده دانسته و نتیجه آن –که عواطف غیرعقلانی هستند– را نمی‌پذیرد (روزر، ۲۰۱۲، ۱۰۶).

روزر می‌گوید عواطفی وجود دارند که از این تقسیم‌بندی دوتایی فراتر می‌روند؛ عواطفی که شامل سطوح بالاتری از تأمّل و روایت‌گری می‌شوند، مثل پاسخ‌های عاطفی ما نسبت به شخصیت‌های تخیلی یا افراد و رویدادهایی که تماس بی‌واسطه با آنها نداریم. این عواطف واکنش سریع و خودبخودی ما به موقعیتی که در آن قرار داریم نیستند، بلکه فرآیندهایی تأمّلی و خودآگاه در ایجاد آنها نقش دارند. روزر این عواطف را *عواطف اخلاقی* می‌داند. با اینکه عواطف خطاپذیر هستند، اما مثلِ خروجی‌های دیگر قوای شناختی ما، عواطف نیز قابل ارزیابی هستند و حتی خود عواطف می‌توانند منبع تأمّلات نقادانه ما درباره عواطف مربوط به خطر باشند. مثلاً همدلی می‌تواند عواطف خودخواهانه فرد را تصحیح کند (روزر، ۲۰۱۲، ۱۰۶–۱۰۷ و ۱۱۰).

بر این اساس، روزر معتقد است مهندسان باید بیاموزند در طراحی مصنوعات از عواطف خود به‌نحوی درست بهره ببرند تا بتوانند تخمین بهتری نسبت به خطرات و آسیب‌های احتمالی مصنوعاتی که تولید می‌کنند، داشته‌باشند. دخالت دادن عواطف در فرآیند طراحی به طراحان اجازه می‌دهد با موقعیت‌ها درگیر شوند، به جای آن که صرفاً با تکیه بر محاسبات و در موضعی منزوی و انتزاعی به مسائل فنی بنگرند (روزر، ۲۰۱۲، ۱۱۱). اگر بخواهیم از دیدگاه روزر درباره هوش مصنوعی استنباطی داشته‌باشیم، می‌توان گفت هر عاملی که بخواهد به طراحی مصنوعات دست بزند (چه انسان و چه هوش مصنوعی) برای برآورد خطر به عاطفه‌مندی نیاز دارد.

۴-۳. عواطف و مجازات‌پذیری

مجازات‌پذیری (یا پاداش‌پذیری) از دو جهت با ظرفیت‌های اخلاقی لازم برای مراتب بالای عاملیت اخلاقی، مرتبط می‌شود. از یک سو، اغلب فرض می‌شود که بین عاملیت اخلاقی و مسئولیت‌پذیری همبستگی وجود دارد: عامل اخلاقی نسبت به عملی که از آن سر می‌زند مسئولیت دارد و کسی را می‌توانیم بابت کاری مسئول بدانیم که در آن کار عاملیت داشته‌است. از سوی دیگر در تفسیر معنای مسئولیت‌پذیری، به مجازات‌پذیری اشاره می‌شود؛ یعنی هنگامی می‌توان عاملی را بابت عملی مسئول دانست که بتوان به نحو معقولی آن را بابت غیراخلاقی بودن آن عمل مجازات کرد (و بالعکس؛ بابت اخلاقی بودن پاداش داد). پس با فرض همبستگی بین عاملیت و مسئولیت‌پذیری و اینکه مسئولیت‌پذیری شامل مجازات‌پذیری می‌شود، می‌توان گفت مجازات‌پذیری، با عاملیت اخلاقی مرتبط و ملازم است (هرچند شرط ضروری برای آن نباشد).

از طرف دیگر، نظریه‌هایی که به چگونگی پیدایش ظرفیت‌های اخلاقی در یک عامل می‌پردازند، به مجازات به‌عنوان یکی از سازوکارهای سوق دادن افراد به سمت رعایت قواعد اخلاقی اشاره می‌کنند. این نظریه‌ها اغلب شیوه شکل‌گیری اخلاق در انسان را (در مقیاس فردی یا اجتماعی) مبنا قرار می‌دهند. رویکردهای رشدی، روندِ شکل‌گیری ظرفیت‌های اخلاقی در هر فرد را در نظر می‌گیرند. مثلا طبق نظریه کولبرگ، کودکان در مراحل اولیه رشد اخلاقی بر اساس پاداش و مجازات‌هایی که از محیط پیرامونشان دریافت می‌کنند، به درکی ابتدایی از درست و غلط بودن اعمال می‌رسند. رویکردهای تکاملی نیز با استفاده از نظریه بازی، نشان می‌دهند مجازات افرادی که مرتکب تخلف اخلاقی می‌شوند (مثلاً تقلبی که منجر به از دست رفتن اعتماد می‌شود) و پاداش دادن به افرادی که مطابق قواعد اخلاقی رفتار می‌کنند، بخشی از سازوکار ایجاد اخلاق در جوامع انسانی است (استورس‌هال، ۲۰۰۷، ۲۹۴-۲۹۵).

بر این اساس، مجازات‌پذیری می‌تواند عاملی مرتبط با وجود یا پیدایش ظرفیت اخلاقی باشد. از سویی دیگر بسیاری از فیلسوفان معتقدند شرط لازم برای مجازات‌پذیری، وجود عواطف است. هویتی مجازات‌پذیر است که بتواند تجربه‌ای از درد و رنج داشته‌باشد. مجازات (و تشویق) هویتی که فاقد تجربه درد (و لذت) باشد بی‌معنا است. ولیز یکی از مدافعین این دیدگاه است. او داشتن حس‌مندی[۱]، تجربه آگاهانه از لذت، رنج و همدلی را شرط لازم برای عاملیت اخلاقی می‌داند (ولیز، ۲۰۲۱، ۴۸۹). او می‌گوید برای داشتن درکی اولیه از ایده اساسی خوب بودن، باید بتوان حس کرد چه چیزی منجر به لذت، معناداری و رضایت می‌شود. هوش مصنوعی نمی‌تواند میل، ترس و امید داشته‌باشد. در نتیجه، محرومیت یا برخورداری برای آن معنا ندارد (ولیز، ۲۰۲۱، ۴۹۳). ولیز مجازات ربات‌ها (مثلاً محبوس کردن یک ربات برای فکر کردن به کارهای ناپسندش!) را یک صحنه‌سازی و در حد نمایشی جالب می‌بیند، اما از دید او این به هیچ وجه یک تنبیه واقعی نیست. چون نمی‌توان هویتی را که چیزی را تجربه یا ارزش‌گذاری نمی‌کند، تنبیه کرد (ولیز، ۲۰۲۱، ۴۹۵).

۴. عواطف به مثابه عاملی منفی در ارتقاء ظرفیت‌های اخلاقی

برخلاف محققینی که نسبتی مثبت بین عاطفه‌مندی مصنوعات و ظرفیت اخلاقی آنها می‌بینند، برخی معتقدند وجود عواطف مانعی برای اخلاقی بودن مصنوعات است. در این بخش چهار دیدگاه در دفاع از این موضع ارائه می‌شود: استدلال‌های «تصمیم‌گیری

---

[۱] Sentience

اخلاقی و رایش عاطفی»، «سراب عواطف»، «پارادوکس انسان‌انگاری و انسانیت‌زدایی» و «مهارت‌زدایی اخلاقی از انسان». در استدلال اول نقش عواطف در نسبت با عاملیت اخلاقی مصنوعات تحلیل می‌شود و ادعا می‌شود برای توسعه ظرفیت اخلاقی مصنوعات، الگو قرار دادن مدل انسانی مناسب نیست. در استدلال‌های دوم و سوم و چهارم، نسبت عواطف و ظرفیت اخلاقی، با یک واسطه به عاملیت اخلاقی مصنوع مرتبط می‌شود. در واقع در این سه استدلال فرض می‌شود روی دیگر عاملیت اخلاقی، مخاطب عمل اخلاقی بودن است و اگر هوش مصنوعی درجاتی از خودمختاری و قصدمندی را بروز دهد، می‌تواند به جایگاه مخاطب عمل اخلاقی نیز ارتقاء پیدا کند. این سه استدلال ابعاد اخلاقی عاطفه‌مندی مصنوع به‌عنوان مخاطب عمل اخلاقی را تحلیل می‌کنند.

۱-۴. تصمیم‌گیری اخلاقی و رایش عاطفی

برخی فیلسوفان معتقدند در الگوی شناختی و عملکردی انسان، عواطف مخلّ تفکر و عمل اخلاقی هستند. از دید این فیلسوفان برای توسعه ظرفیت‌های اخلاقی در هوش مصنوعی، نباید مدل انسانی را الگو قرار دهیم. بنابراین دیگر مسئله پیاده‌سازی عواطفی مشابه انسان در هوش مصنوعی مطرح نیست.

علاوه بر رویکردهای فلسفی (مانند اخلاق کانتی) که استدلال اخلاقی را یک استدلال عقلانی ناب می‌دانند، پژوهش‌های تجربی درباره استدلال و تصمیم‌گیری اخلاقی نیز نشان می‌دهد ورود عواطف به فرآیند تأمّل اخلاقی، می‌تواند نتایج نامقبولی داشته‌باشد. طبق این پژوهش‌ها مثلاً در مورد معروف تنگنای اخلاقی سوزنبان (که در آن افراد در جایگاه سوزنبانی هستند که می‌بینند روی ریل قطار افرادی بسته شده و در معرض مرگ هستند، اما می‌تواند با نوعی مداخله، موجب مرگ افراد کمتری شود)، این که فرد با اهرم مسیر قطار را عوض کند یا با تماس مستقیم کسی را روی ریل بیندازد بر انتخاب افراد اثر می‌گذارد؛ اگر قرار بر لمس کردن قربانی باشد، افراد بیشتر در مقابل این گزینه مقاومت می‌کنند (والاچ و الن، ۲۰۰۹، ۲۰۱). مختل شدن تفکر و تصمیم اخلاقی توسط عواطف با نام «رایش عاطفی»[۱] شناخته می‌شود.

مصنوعات فاقد عواطفی همچون انسان هستند و به این ترتیب نسبت به تهدید رایش اخلاقی ایمن هستند (والاچ و الن، ۲۰۰۹، ۱۴۹). والاچ و الن در توصیف وجه تاریک عاطفه‌مندی ربات‌ها به این نکته اشاره می‌کنند که گاهی هوشمندی متفاوت هوش مصنوعی، می‌تواند به آن نسبت به انسان برتری ببخشد. همانطور که هوش مصنوعی دیپ بلو ۲[۲]به دلیل این که با شیوه‌ای متفاوت از شیوه انسانی شطرنج بازی می‌کرد کاسپاروف را شکست داد، ممکن است عامل اخلاقی هوش مصنوعی هم با ابزارهای شناختی یا عاطفی متفاوت از انسان بتواند عملکرد اخلاقی بهتری داشته‌باشد (والاچ و الن، ۲۰۰۹، ۱۴۲).

باتکس نیز مدافع چنین ایده‌ای است. او پیشنهاد می‌کند اگر بخواهیم از برخی نقائص اخلاقی انسانی اجتناب کنیم، باید هوش مصنوعی طراحی کنیم که شبیه به ما نباشد (باتکس، ۲۰۲۰، ۵). این دیدگاه توسط استورس‌هال با تفصیل و جدیت بیشتری شرح داده شده‌است. استورس‌هال با اشاره به تقسیم‌بندی سه‌گانه نهاد، خود و فراخود[۳]فرویدی، می‌گوید عواطف و احساسات در قلمرو نهاد هستند که نقش اندکی در اخلاقی بودن دارد. در واقع رانه‌های خودخواهی و منفعت‌طلبی که در نهاد وجود دارد، همان چیزی است

---

[۱] Emotional Hijacking
[۲] Deep Blue II
[۳] Id, Ego, Super Ego

که اخلاق باید علیه آن عمل کند (استورس‌هال، ۲۰۰۷، ۳۲۵). سم آلتمن، مدیرعامل و هم‌بنیان‌گذار شرکت اپن‌ای‌آی نیز گفته‌است می‌توانیم سامانه‌های جی.پی.تی بسازیم که سوگیری کمتری از انسان‌ها داشته‌باشند، چون فاقد بار عاطفی هستند (والر، ۲۰۲۴، ۱۵۲).

استورس‌هال رویکردی تکاملی نسبت به توسعه هوش مصنوعی دارد و پیش‌بینی می‌کند سازوکارهایی مشابه تکامل –اما بسیار سریع‌تر– مسیر پیشرفت هوش مصنوعی را هدایت کنند. در مورد انسان‌ها، برخی عواطف مثل خشم در تنظیم روابط اجتماعی نقش داشته‌اند، اما می‌توان آنها را در هوش مصنوعی به‌نحوی بهتر مثلاً با استفاده از پلیس جایگزین کرد (استورس‌هال، ۲۰۰۷، ۳۴۵). استورس‌هال معتقد است با این کار می‌توانیم اخلاقی فراتر از اخلاق انسانی داشته باشیم، هوش مصنوعی می‌تواند راهنمای اخلاقی انسان باشد و به‌واسطه وجود آنها استانداردهای اخلاقی انسانی نیز بالاتر برود (استورس‌هال، ۲۰۰۷، ۳۵۴-۳۵۳).

۲-۴. سراب عواطف

ماساهیرو موری[۱] برای توصیف نمودار میزان راحتی انسان در برخورد با ربات بر اساس میزان شباهت ربات به انسان، از اصطلاح «درّه غریب»[۲] استفاده می‌کند. این اصطلاح به افتی درّه‌مانند در این نمودار اشاره می‌کند که نشان می‌دهد با بیشتر شدن شباهت ربات‌ها به انسان، راحتی انسان‌ها در مواجهه با ربات بیشتر می‌شود تا جایی که این شباهت از آستانه‌ای عبور می‌کند. در اینجا راحتی انسان در برخورد با ربات افت می‌کند و افراد حسی غریب پیدا می‌کنند. اگر ربات‌ها از این محدوده شباهت عبور کنند و بسیار بسیار شبیه به انسان باشند، مجدداً میزان راحتی انسان‌ها با آنها افزایش پیدا می‌کند. شبیه‌سازی ابرازهای عاطفی در ربات، در ایجاد شباهت ربات به انسان و حس اعتمادبرانگیزی نقشی اساسی دارد.

در مواردی که انسان با ربات فعالیتی گروهی را پیش می‌برند، این که ربات ابرازهایی عاطفی از خود نشان دهد (مثلاً هنگامی که می‌خواهد به انسان بگوید سریع‌تر کار کند، با صدایی لرزان و مضطرب سخن بگوید)، باعث می‌شود انسان تمایل بیشتری داشته‌باشد که ربات‌ها عاملیت و سطحی از خودمختاری و هدفمندی داشته‌باشند (شوتز و کرول، ۲۰۰۷، ۵)، اما پیاده‌سازی ابرازهای عاطفی در ربات‌ها دارای شائبه فریبندگی و به‌لحاظ اخلاقی سؤال‌برانگیز است. فاریسکو و همکارانش نیز معتقدند درعین‌حال که فقدان درگیر عاطفی در هوش مصنوعی، ظرفیت اخلاقی آن را کم کرده‌است، اما پیاده‌سازی ظاهری عواطف در هوش مصنوعی نیز می‌تواند فریب‌آمیز باشد (فاریسکو، ۲۰۲۰، ۲۴۲۱).

استورس‌هال نیز که از اساس فقدان عواطف را برگ برنده اخلاقی هوش مصنوعی می‌دانست، معتقد است پیاده‌سازی ظواهر اخلاقی در هوش مصنوعی خطرناک است چرا که می‌توان ظهورات بیرونی عواطف را عیناً در ربات شبیه‌سازی کرد، طوری که تأثیر مد نظرشان را روی انسان‌ها بگذارند. حتی ممکن است ابراز عواطف تصنعی ربات‌ها به حدی برسد که بتوانند همچون انسان‌ها گزارشی از کیفیات پدیداری عواطفشان نیز ارائه دهند. درحالی‌که پشت این ظاهر، فهم یا عواطفی واقعی وجود ندارد. استورس‌هال وضعیت چنین ربات‌هایی را که به خوبی آموخته‌اند در چه موقعیتی چه احساساتی را بروز دهند بی‌آن که آن را حقیقتاً تجربه کنند، مشابه وضعیت ابر–سیاستمداران (احتمالاً بدکردار) و بیماران روانی می‌داند (استورس‌هال، ۲۰۰۷، ۲۹۱-۲۹۲).

---

[۱] Masahiro Mori

[۲] Uncanny Valley

۳-۴. پارادوکس انسان‌انگاری و انسانیت‌زدایی

یکی از پیامدهای پیاده‌سازی عواطف در هوش مصنوعی، تقویت انسان‌انگاری[۱]آن توسط افرادی است که با آن در تعاملند. مشاهده ابرازات عاطفی شبه‌انسانی در هوش مصنوعی باعث می‌شود افراد هرچه بیشتر تمایل داشته‌باشند هوش مصنوعی را مانند انسان درک کرده و با اوصافی انسانی توصیف کنند. درحالی‌که آنها بر خلاف انسان‌ها، به دست انسان‌ها *طراحی* و *ساخته* شده، *خریدوفروش* می‌شوند و همواره *باید* در خدمت رفع نیازهای انسانی باشند. این بخشِ واقعیت، از آنها انسانیت‌زدایی[۲]می‌کند. به این ترتیب عاطفه‌مندی هوش مصنوعی که به تشدید انسان‌انگاری منتهی می‌شود، وضعیت تناقض‌آمیزی را ایجاد می‌کند که کاپوچو و همکاران آن را «پارادوکس انسان‌انگاری و انسانیت‌زدایی هم‌زمان[۳]» می‌خوانند (کاپوچو و همکاران، ۲۰۱۹، ۲۶)، اما چه چیز درباره این وضعیت مشکل‌ساز است؟

گذشته از جنبه تناقض‌آمیز عاطفی و فکری، برخی محققین این وضعیت را به‌لحاظ اخلاقی فسادآور می‌دانند. در ادوار گوناگون زندگی بشری، هویاتی در جایگاه «انسان‌وارِ انسانیت‌زدایی‌شده» قرار گرفته‌اند؛ هویاتی همچون بردگان، یا زنان و کودکان. دیگر افراد جامعه در تعامل با این افراد بخشی از سنجه‌های رفتاری را بر اساس شأنیت انسانی و برخی دیگر را بر اساس شیءانگاری و ابزارانگاری این هویات تعریف می‌کردند. مسئله فسادبرانگیزی که در ارتباط با این پارادوکس می‌تواند مطرح باشد، تعریف و به‌رسمیت شمردن چنین جایگاهی در تعاملات اجتماعی است. در طول تاریخ و در جوامع مختلف، تعریف این جایگاه دستمایه برتری‌طلبی یک گروه اجتماعی نسبت به گروه‌های دیگر و بهره‌کشی و استثمار ایشان بوده‌است. خصوصاً هرچه کفه انسان‌انگاری در این این رابطه قوی‌تر باشد (هرچه گروه فرودست شباهت بیشتری به گروه فرادست داشته‌باشند)، گروه فرادست با سلطه بر ایشان، قدرت خویش را بیشتر به رخ می‌کشد.

عاطفه‌مندی هوش مصنوعی می‌تواند رابطه هوش مصنوعی و انسان را بیشتر از رابطه ابزار و کاربر، به رابطه برده و ارباب شبیه کند. درعین‌حال، برخی ملاحظاتی که در رابطه بین برده و ارباب وجود دارد در این رابطه وجود ندارد. کاربران نیازی به احساس مسئولیت یا پاسخگویی نسبت به هوش مصنوعی نمی‌بینند و دلیلی برای پرهیز از بدرفتاری با آن ندارند. آنها می‌دانند هرگاه بخواهند می‌توانند مصاحبِ انسان‌وار خود را خاموش کنند یا دور بیندازند؛ بی‌آن که این کار از نظر اخلاقی مشکلی داشته‌باشد. از نظر فیلسوفان فضیلت‌گرا این وضعیت موجب تقویت خوی سلطه‌گری بر هم‌نوع در کاربران هوش مصنوعی شده و زمینه‌ساز رشد رذایل اخلاقی در ایشان می‌شود.

۴-۴. مهارت‌زدایی اخلاقی از انسان

عاطفه‌مندی باعث اجتماعی‌تر شدن هوش مصنوعی و تحکیم جایگاه آن در روابط اجتماعی انسان‌ها می‌شود. برخی محققین نسبت به این روند نگرانی‌هایی را ابراز می‌کنند. از دید ایشان، با افزایش ارتباطات انسان–ماشین و غنی‌تر شدن این روابط، به تدریج این ارتباطات جای روابط انسان–انسان را می‌گیرد و منجر به کاهش روابط بین انسانی می‌شود. این وضعیت مسائل مختلفی را مطرح می‌کند، از جمله تأثیر جایگزینی روابط انسان–ماشین بر مهارت‌های اخلاقی انسان‌ها.

---

[۱] Anthropomorphism

[۲] Dehumanization

[3] Anthropomorphizing while Dehumanizing Robots Paradox- ADP

ونگ با اشاره به تحقیقات تجربی که نشان می‌دهند کار با دستیاران هوشمند مانند الکسا یا سیری باعث می‌شود افراد و خصوصاً کودکان گستاخ شوند، ادعا می‌کند از آنجا که دستیار دیجیتال بدون ملاحظه شیوه رفتار کاربر به او پاسخ می‌دهد، حسی کاذب از شایستگی در فرد ایجاد می‌کند. به عقیده او کسب فضایل اخلاقی با ممارست و دریافت بازخوردهای مناسب حاصل می‌شود و اگر فناوری فرصت‌های کمتری به افراد برای تمرین مهارت‌های اخلاقی بدهد، یا باعث شود بازخوردهای مناسبی دریافت نکنند، افراد به‌لحاظ اخلاقی مهارت‌زدایی[1] می‌شوند (ونگ، ۲۰۱۹، ۲).

والر در کتاب *آینه هوش مصنوعی*[2] با تفصیل بیشتر به این مسئله می‌پردازد. والر که نام کتابش را از این بیت مولوی الهام گرفته‌است که می‌گوید «آینه‌ت دانی چرا غمّاز نیست؟ زانکه زنگار از رخش ممتاز نیست»، ادعا می‌کند ما در فناوری‌ها تصویری از خود را جست‌وجو می‌کنیم، اما هوش مصنوعی آینه‌ای زنگارگرفته است که تصویری واقع‌نمایانه از ما به نمایش نمی‌گذارد. او به طور خاص فناوری‌های هوش مصنوعی را که با هدف شبیه‌سازی روابط عاطفی توسعه یافته‌اند مانند چت‌بات‌ها و برنامه‌های کاربردی مثل شیائوآیس[3] و رپلیکا[4] مورد بررسی قرار می‌دهد. از دید او این فناوری‌ها به دنبال این هستند که مشکلی را که ناشی از بیگانه‌شدگی انسان در عصر فناوری است، با راه حلی فناورانه مرتفع سازند، درحالی‌که صرفاً سایه‌هایی غریب از معاشرت‌های انسانی را ایجاد کرده و تصویری وهمی از یک رابطه عاطفی متقابل و کامل را ارائه می‌دهند (والُر، ۲۰۲۴، ۱۴۴).

در برنامه‌های هوش مصنوعی که با هدف تأسیس رابطه عاطفی ساخته شده‌اند، کاربرانی که امکان معاشرت‌های انسانی باکیفیت برایشان فراهم نیست، می‌توانند با هوش مصنوعی که مطابق سلیقه‌شان اختصاصی‌شده رابطه‌ای را که می‌پسندند (افلاطونی، رمانتیک، یا جنسی) برقرار کنند. بسیاری از کاربران این برنامه‌ها چتبات رمانتیک خود را شریک عاطفی، عاشق، یا حتی همسر به حساب می‌آورند. در این روابط، چت‌بات‌ها همان چیزی را بروز می‌دهند که مطابق میل کاربر است (حتی اگر خواسته کاربر، شخصیتی سرد و بی‌تفاوت باشد). به همین دلیل کاربران رابطه با این چتبات‌ها را امن‌تر و برتر از رابطه با انسان‌ها می‌دانند و بعضاً تجربه خود را احساس نوعی عشق بی‌قیدوشرط توصیف می‌کنند (والُر، ۲۰۲۴، ۱۴۴–۱۴۵).

والر معتقد است این آینه‌های کدر، تصویری تحریف‌شده از عشق – به‌عنوان یک فضیلت اخلاقی– ارائه می‌دهند. عشق واقعی، آنچه ممیزه نوع بشر است و او را به‌لحاظ اخلاقی ارتقاء می‌دهد، همواره سهل‌الوصول، دوطرفه، رضایت‌بخش یا لذت‌بخش نیست. عشق هزینه دارد؛ می‌تواند سخت، خسته‌کننده و دردناک باشد و همواره خطرپذیری و فقدان را با خود به همراه دارد. تصویری که چت‌بات‌های عاطفی از عشق ارائه می‌دهند یک لذت عاشقانه تضمین‌شده است. درحالی‌که در جهان واقعی، عشق تضمینی ندارد. فردی که در رابطه با چت‌بات است می‌داند که اگر بد رفتار کند با او خوب رفتار خواهدشد و اگر خوب رفتار کند با او بهتر رفتار خواهدشد. می‌داند که هرگز به او خیانت نمی‌شود و همه چیز تحت کنترل او باقی خواهدماند، اما این تصویر دروغین است و وابسته شدن افراد به آن، آنها را از درگیر شدن در روابط انسانی واقعی – که در آن عشق می‌تواند بستر رشد فضایل اخلاقی باشد– ناتوان می‌کند. چنان که برای مثال پس از آنکه در نسخه بروزرسانی‌شده رپلیکا، استفاده از عبارات تحریک‌کننده محدود شد؛ افرادی که برای برآورده کردن نیازهای عاطفی–جنسی خود بر این برنامه متکی بودند، گزارش‌هایی از افسردگی، اضطراب و حتی افکار خودکشی

[1] Deskilling
[2] AI Mirror
[3] Xiaoice
[4] Replika

ارائه دادند (والُر، ۲۰۲۴، ۱۴۶–۱۴۸). بر این اساس، والر ادعا می‌کند هوش مصنوعی عاطفه‌مند با ارائه تصویری تحریف‌شده از حقیقت، موجب تضعیف معرفت ما نسبت به خودمان شده و از این طریق ما را از فرصت شکل‌دهی به تجربیات اخلاقی غنی محروم می‌کند (والُر، ۲۰۲۴، ۱۵۲).

۵. نقاط نزاع و چالش‌های پیش رو

جهت‌گیری متفاوت استدلال‌های فوق به‌خوبی نشان‌دهنده چیزی هستند که والاچ و الن آن را کیفیت پارادوکسیکال طراحی مصنوعات عاطفه‌مند می‌نامند (والاچ و الن، ۲۰۰۹، ۱۴۲). برای تحلیل این وضعیت مناقشه‌برانگیز، توجه به پیش‌فرض‌های استدلال‌ها سودمند است. به نظر می‌رسد دو پیش‌فرض در تحلیل نقش عواطف در ظرفیت اخلاقی هوش مصنوعی مؤثرند: لحاظ کردن جنبه پدیداری یا کارکردی عواطف و نقش سازنده یا مخرّبِ عواطف انسانی در اخلاقی بودن. در جدول زیر می‌توان استدلال‌های فوق را از این جهات دسته‌بندی کرد:

استدلال	خصلت عواطف	نقش عواطف انسانی در اخلاقی بودن	تأثیر عواطف بر ظرفیت اخلاقی مصنوعات
حساسیت اخلاقی	پدیداری/ کارکردی	سازنده	مثبت
عقلانیت محدود	پدیداری	سازنده	مثبت
برآورد خطر	پدیداری	سازنده	مثبت
مجازات‌پذیری	پدیداری	سازنده	مثبت
ربایش عاطفی	پدیداری	مخرب	منفی
سراب عواطف	کارکردی	سازنده	منفی
پارادوکس انسان‌انگاری و انسانیت‌زدایی	کارکردی	سازنده	منفی
مهارت‌زدایی اخلاقی	کارکردی	سازنده	منفی

جدول ۱. مفروضات استدلال‌ها درباره عواطف در انسان و مصنوعات

در استدلال‌های فوق (به جز استدلال ربایش عاطفی)، انسانِ اخلاقی به‌عنوان الگوی ساخت مصنوعات اخلاقی در نظر گرفته‌شده و نقش عواطف نیز در اخلاقی بودن انسان‌ها مثبت لحاظ شده‌است. از سویی دیگر، عواطف در انسان‌ها هم جنبه‌ای پدیداری و هم جنبه‌ای کارکردی دارند. جنبه پدیداری عواطف، احساس، تجربه، گرایش[۱] و باورهای خاصی را در انسان ایجاد می‌کند. درحالی‌که جنبه کارکردی به اعمال و تغییرات مشهود منتهی می‌شود. در مورد مصنوعات، عواطف صرفاً به صورت کارکردی قابل پیاده‌سازی هستند. در اغلب استدلال‌های فوق (به جز استدلال ربایش عاطفی و استدلال حساسیت اخلاقی، مورد گویندراجولو و همکاران) آنجا که به نقش مثبت عواطف در ظرفیت اخلاقی اشاره می‌شود، جنبه پدیداری عواطف لحاظ می‌شود. این نقش پدیداری برای برخی به‌اندازه‌ای

Disposition

اهمیت دارد که به دلیل ناممکن دانستن وجه پدیداری عواطف در مصنوعات، آنها را فاقد عاملیت اخلاقی، یا برای نقش‌هایی با ظرفیت بالای اخلاقی نامناسب می‌دانند.

تقریباً تمام محققین می‌پذیرند که جنبه‌های کارکردی عواطف –علی‌الاصول– تا حد زیادی قابل پیاده‌سازی در هوش مصنوعی هستند، اما در سه استدلالِ سراب عواطف، پارادوکس انسانیت‌انگاری و انسانیت‌زدایی و مهارت‌زدایی، از یک سو نقش عواطف در اخلاقی بودن انسان‌ها سازنده در نظر گرفته‌شده‌است و از سویی دیگر پیاده‌سازی صرفِ جنبه کارکردی عواطف در هوش مصنوعی به‌لحاظ پیامدهای مربوط به ظرفیت اخلاقی، منفی شناخته شده‌است. می‌توان گفت هر سه این استدلال‌ها بر *فریبندگی* و *تقلبی بودن* عواطف مصنوعی دلالت دارند؛ گویی عواطف مصنوعی هرگز نمی‌توانند یک عاطفه کامل باشند و شکاف عاطفی میان انسان و فناوری پُرشدنی نیست. در این صورت، هرچه بیشتر ظاهر آنها به ظاهر عواطف انسانی شبیه‌تر باشد، فریب بزرگ‌تری رخ داده‌است و پیامدهای اخلاقی منفی آن شدیدتر خواهدبود. قابل ذکر است که استدلال حساسیت اخلاقی گویندا‌راجولو و همکارانش که پیاده‌سازی عواطف را در راستای تشخیص الگوی اخلاقی ممکن می‌دیدند، نتیجه‌ای برخلاف استدلال‌های فوق دارد، اما حتی اگر پروژه ساخت ربات فضیلت‌مند را پروژه‌ای امکان‌پذیر بدانیم، همچنان خطرات گوناگونی که استدلال‌های سه‌گانه فوق به آن دلالت دارند می‌تواند دلایلی قوی برای منفی شمردن تأثیر پیاده‌سازی ظاهری عواطف در هوش مصنوعی به دست دهد.

به این ترتیب، می‌توان گفت علی‌رغم مناقشات ظاهری که درباره نقش عواطف در ظرفیت اخلاقی مصنوعات وجود دارد؛ با تفکیک میان جنبه پدیداری و کارکردی عواطف و توجه به این که تنها جنبه کارکردی عواطف در هوش مصنوعی قابل پیاده‌سازی هستند؛ درباره پیامدهای اخلاقی منفی پیاده‌سازی عواطف در هوش مصنوعی، اتفاقی نظری تقریبی وجود دارد، اما این نتیجه را می‌توان با یک پرسش معرفت‌شناختی به چالش کشید: بر چه اساس می‌توان از فقدان وجه پدیداری عواطف در هوش مصنوعی اطمینان داشت؟ به عبارت دقیق‌تر، با توجه به اینکه ما وجود عواطف پدیداری در انسان‌های دیگر را از ظواهر و رفتارهایشان (جنبه‌های کارکردی عواطف) می‌فهمیم، چه ملاکی برای تشخیص وجه پدیداری عواطف می‌توان ارائه داد که وجود آن را در انسان‌های دیگر تأیید کند؛ اما درباره هوش مصنوعی نه؟

در پاسخ به این سؤال و برای تعیین این ملاک، نمی‌توان به هیچ یک از ویژگی‌های ظاهری، رفتاری و کارکردی ارجاع داد؛ چرا که طبق نظر متخصصین در چشم‌انداز توسعه هوش مصنوعی کاملاً محتمل است که هوش مصنوعی از تمامی این جهات از انسان‌ها تقریباً غیرقابل شناسایی بشود، اما ملاک دیگری وجود دارد که در عین حال که محدود به زاویه دید اول‌شخص نیست و وجود وجه پدیداری عواطف را در *اذهان دیگر* شناسایی می‌کند، اما هوش مصنوعی را واجد این ویژگی نمی‌داند. این ملاک، به *تاریخچه علّی پیدایش* ارجاع می‌دهد. ما به ذهن افراد دیگر دسترسی اول‌شخص نداریم و نمی‌توانیم (آن‌طور که به ذهن خود دسترسی داریم) کیفیات پدیداری آنها مستقیماً دسترسی داشته باشیم. با این حال، از منظری سوم‌شخص می‌توانیم ببینیم که تاریخچه علّی شکل‌گیری ما و انسان‌های دیگر شباهت‌های بسیاری با هم دارد. فرآیند زیستی که منتهی به پیدایش انسان‌ها می‌شود بسیار شبیه به یکدیگر است، اما با فرآیند طراحی و ساخت که منتهی به پیدایش مصنوعات می‌شود بسیار متفاوت است. بنابراین، با تکیه بر قرائن مربوط به شباهت و تفاوت در زنجیره علّی پیدایش، می‌توانیم انسان‌های دیگر را از نظر کیفیات ذهنی شبیه به خود بدانیم و از این نظر از مصنوعات متمایز کنیم.

نتیجه‌گیری

آیا پیاده‌سازی عواطف در هوش مصنوعی به بهبود ظرفیت‌های اخلاقی آن کمک می‌کند؟ برای پاسخ به این پرسش، به تدقیق مفهوم عاملیت اخلاقی پرداختیم. رابطه عواطف با عاملیت اخلاقی در برخی موارد سرراست است و عواطف (چه به طور ایجابی و چه سلبی) به‌عنوان شرایط صریح یا ضمنی عاملیت اخلاقی مطرح هستند، اما در برخی موارد این نسبت غیرمستقیم است.

از میان زوایای مختلف نسبت عواطف و ظرفیت‌های اخلاقی مصنوعات، بر استدلال‌هایی که به نقش مثبت یا منفی عواطف در ظرفیت‌های اخلاقی مصنوعات می‌پرداخت متمرکز شدیم. در چهار استدلال «حساسیت اخلاقی»، «عقلانیت محدود»، «برآورد خطر» و «مجازات‌پذیری»، انسان به‌عنوان الگوی اخلاقی برای ساخت مصنوعات فرض می‌شوند و عواطف دارای نقشی مثبت در ظرفیت‌های اخلاقی شناخته می‌شوند. در چهار استدلال «ربایش عاطفی»، «سراب عواطف»، «پارادوکس انسان‌انگاری و انسانیت‌زدایی همزمان» و «مهارت‌زدایی اخلاقی»، وجود عواطف دارای نقشی منفی در ظرفیت‌های اخلاقی مصنوعات شناخته می‌شود. با تحلیل این استدلال‌ها دیده می‌شود (به جز یک مورد)، دو جنبه پدیداری و کارکردی عواطف در انسان در کنار یکدیگر نقشی سازنده در اخلاقی بودن انسان دارند، اما از آنجا که در پیاده‌سازی عواطف در هوش مصنوعی تنها جنبه‌های کارکردی آن قابل پیاده‌سازی است، شکاف میان عامل اخلاقی مصنوعی و عامل اخلاقی انسانی پر نخواهدشد و ساخت این عواطف با نوعی فریب همراه است. این فریب از جهات مختلفی می‌تواند بر شخصیت اخلاقی انسان‌هایی که با این ربات‌ها در تعامل هستند، پیامدهایی منفی به جای بگذارد.

References

Behdadi, D. & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency, *Minds and Machines*, 30 (2), 195-218. https://doi.org/10.1007/s11023-020-09525-8

Brey, P., (2014). From Moral Agents to Moral Factors: The Structural Ethics Approach, in *The Moral Status of Technical Artefacts*, Eds. P. Kroes & P. P. Verbeek. Springer.

Bringsjord, S. (2007). Ethical Robots: The Future Can Heed Us, *AI and Society*, 22 (4),539-550. https://doi.org/10.1007/s00146-007-0090-9

Butkus, M. A. (2020). The Human Side of Artificial Intelligence, *Science and Engineering Ethics*, 26 (5), 2427-2437. https://doi.org/10.1007/s11948-020-00239-9

Cappuccio, M.; Peeters, A. & McDonald, W. (2019). Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition, *Philosophy & Technology*, 33 (1), 9-31. https://doi.org/10.1007/s13347-019-0341-y

Coeckellbergh, M., (2007). Violent Computer Games, Empathy, and Cosmopolitanism, *Ethics and Information Technology*, 9 (3), 219-231. https://doi.org/10.1007/s10676-007-9145-3

Farisco, M.; Evers, K. & Salles, A. (2020). Towards Establishing Criteria for the Ethical Analysis of Artificial Intelligence, *Science and Engineering Ethics*, 26 (5), 2413-2425. https://doi.org/10.1007/s11948020-00238-w

Floridi, L. & Sanders J. W. (2004). On the Morality of Artificial Agents, *Minds and Machines*, 14 (3), 349-379. https://doi.org/10.1023/B: MIND.0000035461.63578.9d

Franklin S. & Graesser A. (1996). Is it an Agent, or just a Program? A Taxonomy for Autonomous Agents, *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, Springer-Verlag.

Govindarajulu N. S.; Bringsjord S.; Ghosh R. & Vasanth S. (2019). Toward the Engineering of Virtuous Machines. In *AAAI/ACM Conference on AI, Ethics, and Society* (AIES '19), January 27–28, 2019, Honolulu.

Himma, K. E. (2009). Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent? *Ethics and Technology*. 11 (1), 19-29. https://doi.org/10.1007/s10676-008-9167-5

Johnson, D. G. & Noorman M. (2014). Artefactual Agency and Artefactual Moral Agency, in *The Moral Status of Technical Artefacts*, Eds. P. Kroes & P. P. Verbeek, Springer.

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21(4),18-21. https://doi.org/10.1109/MIS.2006.80

Nyholm, S. (2019). Other Minds, Other Intelligences: The Problem of Attributing Agency to Machines, *Cambridge Quarterly of Healthcare Ethics*, 28 (4), 592-598. https://doi.org/10.1017/S0963180119000537

Orteny A.; Clore G. & Colins A. (2022). *The Cognitive Structure of Emotions*, Cambridge University Press.

Pitt, J. (2014). Guns Don't Kill, People Kill: Values in/and or around Technologies, in *The Moral Status of Technical Artefacts*, Eds Kroes P. & P. P. Verbeek, Springer.

Powers, T. M. (2013). On the Moral Agency of Computers, *Topoi*, 32 (2), 227-236. https://doi.org/10.1007/s1124512-9149-4

Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*, Oxford University Press.

Roeser, S. (2009). The Relation between Cognition and Affect in Moral Judgments about Risks, in *The Ethics of Technological Risk*, Eds Saved & Roeser, Earthscan Publication Ltd.

Roeser, S. (2010). Emotional Reflection about Risks, in *Emotions and Risky Technologies*, Ed R. Sabine, Springer.

Roeser, S. (2012). Emotional Engineers: Toward Morally Responsible Design, *Science Engineering Ethics*, 18 (1),103-115. https://doi.org/10.1007/s11948-010-9236-0

Scheutz, M. C. & Crowell, C. (2007). The Burden of Embodied Autonomy: Some Reflections on the Social and Ethical Implications of Autonomous Robots. *Paper presented at the Workshop on Roboethics at the International Conference on Robotics and Automation*, Rome.

Simon, H. A. (1967). Motivation and emotional controls of cognition. *Psychological Review*, 74 (1), 29-39. https://doi.org/10.1037/h0024127

Sparrow, R. (2016). Kicking a Robot Dog, *ACM/IEEE 11th International Conference on Human-Robot Interaction*, 229-229.

Vallor, S. (2024). *The AI Mirror: How to Reclaim Our Humanity in the Age of Machine Thinking*, Oxford University Press.

Verbeek, P. P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*, Chicago Press.

Véliz, C. (2021). Moral Zombies: Why Algorithms are not Moral Agents, *AI and Society*, 36 (2), 487-497. https://doi.org/10.1007/s00146-021-01189-x

Winner, L. (1980). Do Artefacts Have Politics? *Daedalus*, 109 (1),121-136.

Wong, P. (2019). Rituals and Machines: A Confucian Response to Technology-Driven Moral Deskilling, *Philosophies*, 4 (4), 59. https://doi.org/10.3390/philosophies4040059