

A Study and Evaluation of the Mindless Morality Theory in Attributing Moral Agency to Artificial Intelligent Agents

Aboulfazl Aboumohammadi Ardakan¹, Mahdi Zakeri²

1. Department of Philosophy, Faculty of Theology, Farabi Campus, University of Tehran, Qom, Iran. E-mail: a.aboumohammadi@ut.ac.ir
2. Corresponding author, Department of Philosophy, Faculty of Theology, Farabi Campus, University of Tehran, Qom, Iran. E-mail: zakeri@ut.ac.ir

Article Info

Article type:

Research Article

Article history:

Received: 2026 January 28

Revised: 2026 May 13

Accepted: 2026 June 17

Published online:

Keywords:

Moral Agency,
Intelligent Artificial Agent,
Mindless Morality,
Luciano Floridi,
Standard Account of Moral
Agency

ABSTRACT

The concept of moral agency is fundamental to ethical theory, and the question of attributing it to intelligent artificial agents is a central topic in both moral philosophy and machine ethics. Luciano Floridi, in his theory of “mindless morality,” critiques the traditional anthropocentric approach and attempts to extend moral agency to artificial agents. Drawing on the foundations of “information ethics” and “information ontology,” he replaces conditions such as consciousness, intentionality, and free will with criteria such as interactivity, autonomy, and adaptability for moral agency. This study first explains the theoretical foundations of Floridi’s view and the concept of the “level of abstraction,” then critically examines it in comparison with the standard account of moral agency. The findings indicate that although the theory of “mindless morality” is significant for expanding the scope of moral consideration to intelligent technologies and analyzing emerging issues, it faces serious philosophical challenges. These challenges arise from reducing moral agency to functional criteria, neglecting the role of mental states, and making moral agency dependent on the chosen level of abstraction. Finally, it is argued that acknowledging the moral role of artificial agents does not necessarily entail granting them moral agency in the strict philosophical sense. The standard account of moral agency, due to the essential connection it posits between agency and mentality, retains greater explanatory coherence and ontological robustness.

Cite this article: Aboumohammadi Ardakan, Aboulfazl., & Zakeri, Mahdi. (2026). A study and evaluation of the mindless morality theory in attributing moral agency to artificial intelligent agents. *Philosophy of Religion*, 58 (X), 1-20. <http://doi.org/00000000000000000000>



© The Author(s).

DOI: <http://doi.org/0000000000000000000000>

Publisher: University of Tehran Press

بررسی و ارزیابی انتقادی نظریه «اخلاق بدون ذهن» در انتساب عاملیت اخلاقی به عامل‌های هوشمند مصنوعی

ابوالفضل ابومحمادی اردکان^۱، و مهدی ذاکری^۲

۱. گروه فلسفه، دانشکده الهیات، دانشکدگان فارابی، دانشگاه تهران، قم، ایران. رایانامه: a.aboumohammadi@ut.ac.ir
۲. نویسنده مسئول، گروه فلسفه، دانشکده الهیات، دانشکدگان فارابی، دانشگاه تهران، قم، ایران. رایانامه: zaker@ut.ac.ir

اطلاعات مقاله	چکیده
نوع مقاله: مقاله پژوهشی	مفهوم عاملیت اخلاقی از مفاهیم بنیادین در نظریه اخلاق است و مسئله انتساب آن به عامل‌های هوشمند مصنوعی از مباحث مهم فلسفه اخلاق و اخلاق ماشین به شمار می‌رود. لوچانو فلوریدی در نظریه «اخلاق بدون ذهن» با نقد رویکرد انسان‌محور سنتی، تلاش می‌کند عاملیت اخلاقی را به عامل‌های مصنوعی نیز تعمیم دهد. او با اتکا به مبانی «اخلاق اطلاعات» و «هستی‌شناسی اطلاعاتی»، معیارهایی همچون تعامل‌پذیری، خودمختاری و سازگاری را جایگزین شروطی مانند آگاهی، قصدیت و اراده آزاد برای عاملیت اخلاقی می‌داند. این پژوهش ابتدا مبانی نظری دیدگاه فلوریدی و مفهوم «سطح انتزاع» را تبیین کرده و سپس آن را در مقایسه با روایت استاندارد عاملیت اخلاقی به صورت انتقادی بررسی می‌کند. نتایج نشان می‌دهند که هرچند نظریه «اخلاق بدون ذهن» در گسترش دامنه توجه اخلاقی به فناوری‌های هوشمند و تحلیل مسائل نوپدید اهمیت دارد، اما به دلیل تقلیل عاملیت اخلاقی به معیارهای کارکردی و نادیده گرفتن نقش حالات ذهنی و وابسته‌سازی عاملیت اخلاقی به سطح انتزاع، با چالش‌های فلسفی جدی مواجه است. در نهایت استدلال می‌شود که پذیرش نقش اخلاقی عامل‌های مصنوعی لزوماً مستلزم پذیرش عاملیت اخلاقی به معنای فلسفی دقیق آن نیست و روایت استاندارد عاملیت اخلاقی، به دلیل پیوند ضروری میان عاملیت و ذهنیت، همچنان از انسجام تبیینی و بنیان هستی‌شناختی قوی‌تری برخوردار است.
تاریخ دریافت: 1404/11/8 تاریخ بازنگری: 1405/2/23 تاریخ پذیرش: 1405/3/27 تاریخ انتشار:	
کلیدواژه‌ها: عاملیت اخلاقی، عامل‌های هوشمند مصنوعی، اخلاق بدون ذهن، روایت استاندارد عاملیت اخلاقی، لوچانو فلوریدی.	
استناد: ابومحمادی اردکان، ابوالفضل؛ و ذاکری، مهدی (1405). بررسی و ارزیابی نظریه «اخلاق بدون ذهن» در انتساب عاملیت اخلاقی به عامل‌های هوشمند مصنوعی. فلسفه دین، 58 (1)، 1-20. http://doi.org/000000000000000000000000	
© نویسندگان.	
ناشر: انتشارات دانشگاه تهران.	



مقدمه

پیشرفت‌های روزافزون و توسعه‌ی سریع علوم رایانه و فناوری هوش مصنوعی، همراه با گسترش کاربردهای آن در جنبه‌های مختلف زندگی، موجب دگرگونی بسیاری از روابط و تعاریف سنتی شده و در مواردی، ساختارهای پیشین را به‌طور بنیادین تحت تأثیر قرار داده است. در نتیجه‌ی این تحولات، بخش قابل توجهی از فرایندهای محاسباتی و حتی برخی سازوکارهای شبه‌شناختی که پیش‌تر به ذهن انسان نسبت داده می‌شدند، به سامانه‌های هوش مصنوعی واگذار شده‌اند. این فناوری، به واسطه‌ی توان محاسباتی و قابلیت‌های الگوریتمی خود، توانسته در حوزه‌های گوناگون به دستاوردهای قابل توجهی برسد و پیش‌بینی می‌شود در آینده‌ای نه‌چندان دور، در برخی حوزه‌ها عملکردهایی هم‌تراز با انسان از خود نشان دهد و تجربه‌ی انسانی را به‌گونه‌ای معنادار دگرگون سازد. افزون بر این، سطح عملکرد این سامانه‌ها به‌گونه‌ای ارتقا یافته است که دیگر نمی‌توان آن‌ها را صرفاً ابزارهای منفعل تلقی کرد، بلکه در بسیاری موارد مستلزم نوعی تعامل و درک متقابل میان انسان و ماشین هستند. از این‌رو، می‌توان گفت با شکلی نوین از فناوری مواجه هستیم که نه صرفاً یک مصنوع فنی، بلکه عاملی پیچیده و اثرگذار در زیست‌جهان انسانی به شمار می‌رود. گسترش استفاده از عامل‌های هوشمند مصنوعی، مرزهای میان این فناوری و عامل‌های انسانی را تا حد زیادی مبهم ساخته و مسائل فلسفی و اخلاقی جدیدی را پدید آورده است.

یکی از مهم‌ترین این مسائل، امکان یا عدم امکان انتساب عاملیت اخلاقی به سامانه‌های هوشمند مصنوعی است؛ مسئله‌ای که در کانون بحث‌های فلسفه‌ی اخلاق و فلسفه‌ی عمل قرار دارد. به‌طور کلی، عامل اخلاقی به موجودی اطلاق می‌شود که توانایی انجام کنش‌هایی را دارد که قابلیت ارزیابی اخلاقی از حیث درستی یا نادرستی، یا فضیلت و رذیلت را دارا باشند (McKenna, 2012, p. 11). این مسئله در حوزه‌ی موسوم به «اخلاق ماشین» مورد بررسی قرار می‌گیرد؛ حوزه‌ای که به تحلیل امکان یا عدم امکان عاملیت اخلاقی در موجودات مصنوعی مانند رایانه‌ها و ربات‌ها تحت عنوان «عاملیت اخلاقی مصنوعی» می‌پردازد. در این زمینه، برخی رویکردهای فلسفی، اجتماعی و فنی، تمایل یافته‌اند سامانه‌های هوش مصنوعی را به‌عنوان عامل‌های اخلاقی در نظر بگیرند و از نوعی گسترش قلمرو عاملیت اخلاقی فراتر از انسان دفاع کنند. بر این اساس، جایگاه عاملیت اخلاقی دیگر به‌صورت انحصاری به انسان محدود نمی‌شود. عامل‌های هوشمند مصنوعی، از جمله ربات‌های انسان‌نما، دستیارهای صوتی و چت‌بات‌ها، رفتارهایی پیچیده و گاه اثرگذار از خود نشان می‌دهند که پیامدهای اخلاقی و حقوقی قابل توجهی به همراه دارد. یکی از برجسته‌ترین دیدگاه‌ها در این زمینه، متعلق به لوچانو فلورییدی^۱، فیلسوف معاصر حوزه اخلاق اطلاعات، است که نظریه‌ی «اخلاق بدون ذهن»^۲ را مطرح کرده است. این نظریه بخشی از چارچوب گسترده‌تر اخلاق اطلاعات در اندیشه‌ی او به شمار می‌رود. او این ایده را ابتدا در مقاله‌ای با عنوان «اخلاق اطلاعات: در مورد مبانی نظری اخلاق کامپیوتر» (۱۹۹۹) منتشر کرد که در آن از اخلاق کامپیوتر به اخلاق اطلاعات روی آورد. او در آثار

¹ Luciano Floridi

² Mindless Morality

مهم دیگر خود نیز به این موضوع پرداخته است. از جمله در این زمینه می‌توان به «اخلاق اطلاعات» (2013) و «درباره‌ی اخلاقی بودن عامل‌های مصنوعی» (2004) اشاره کرد. فلوریدی با ارائه‌ی معیارهایی نو برای عاملیت اخلاقی، تلاش می‌کند نشان دهد که قلمرو اخلاق محدود به انسان نیست و می‌توان برای عامل‌های هوشمند مصنوعی نیز، به‌عنوان کنشگران غیرانسانی، نوعی عاملیت اخلاقی قائل شد. او این دیدگاه را در تقابل با روایت استاندارد عاملیت اخلاقی و در چارچوبی تجدیدنظرطلبانه مطرح می‌کند.

دیدگاه استاندارد عاملیت اخلاقی، هم در اخلاق نظری و هم در اخلاق عملی جایگاهی بنیادین دارد و ردپای آن را می‌توان در سنت فلسفی از دوران باستان تا مکاتب اخلاقی معاصر مشاهده کرد. هر سه نظریه‌ی اصلی اخلاق هنجاری، یعنی فایده‌گرایی (در هر دو صورت کلاسیک و معاصر آن)، وظیفه‌گرایی کانتی و اخلاق فضیلت، به‌نحوی پیش‌فرض‌های این دیدگاه را می‌پذیرند. وجه تمایز این نظریه‌ها نه در انکار عاملیت اخلاقی، بلکه در تبیین محتوای کنش اخلاقی و معیارهای ارزیابی آن است (Coeckelbergh, 2010, p. 213). بر اساس روایت استاندارد، عاملیت اخلاقی مستلزم برخورداری از ویژگی‌هایی همچون آگاهی، قصدیت^۳، ظرفیت‌های شناختی پیچیده و حالات ذهنی است. در این چارچوب، تنها موجوداتی که واجد حالات ذهنی درونی مانند باورها، خواسته‌ها و توانایی تأمل و استدلال اخلاقی هستند، می‌توانند به‌عنوان عامل اخلاقی شناخته شوند (Behdadi & Munthe, 2018, p. 38). بر این اساس، انسان‌ها نمونه‌های پارادایمی عامل اخلاقی محسوب می‌شوند و هرگونه انحراف از این الگوی پارادایمی نیازمند استدلال و توجیهی قوی است.

در ادامه، دیدگاه فلوریدی در این زمینه به‌اختصار تبیین شده و سپس در مقایسه با روایت استاندارد عاملیت اخلاقی مورد ارزیابی انتقادی قرار خواهد گرفت.

تبیین نظریه «اخلاق بدون ذهن»

فلوریدی نظریه «اخلاق بدون ذهن» را در چارچوب رویکرد و نظریه «اخلاق اطلاعات»^۴ ارائه کرده است. او در کتاب‌ها و مقالات گوناگون، توسعه فلسفی و نظام‌مند این نظریه را به عنوان یک موضوع جدید در متافیزیک و نظریه اخلاق کلان خود مطرح ساخته و به تشریح چارچوب‌های اساسی و اصلی آن پرداخته است. نظریه اخلاق اطلاعات او مبتنی بر مؤلفه‌ها، مبانی و اصول هستی‌شناختی خاصی است که ریشه در نگاه کامپیوتری و محاسباتی او به جهان دارد و از آن با عنوان هستی‌شناسی اطلاعاتی^۵ یاد می‌شود.

³ Intentionality

⁴ Information Ethics

⁵ Informational Ontology

دو مؤلفه محوری هستی‌شناسی اطلاعاتی فلوریدی عبارت‌اند از: هستی‌محور بودن^۶ و اطلاعات‌محور بودن^۷. از نظر او، اطلاعات عنصر نهایی واقعیت یا وجود و به مثابه جوهر هستی است. هر هستنده یا موجودیتی در جهان، یک هستنده اطلاعاتی^۸ محسوب می‌شود که هم دارای ساختار اطلاعاتی است و هم می‌توان آن را به صورت اطلاعاتی تعریف کرد. به عبارت دیگر، همه هستنده‌ها و موجودیت‌ها بسته‌های منسجم از اطلاعات یا خوشه‌های اطلاعاتی^۹ هستند که دارای داده‌ها و الگوهای اطلاعاتی‌اند و شأن اطلاعاتی دارند.

تمام این هستنده‌ها و موجودات در گستره‌ای به نام سپهر اطلاعات^{۱۰} قرار دارند و شامل همه موجودات از انسان، حیوانات و گیاهان گرفته تا اشیای بی‌جان، نقاشی‌ها، کتاب‌ها، ستارگان و سنگ‌ها، حتی سیستم‌های اطلاعاتی فاقد ذهن (مانند شبکه‌ها، داده‌ها و موجودات غیرآگاه) می‌شوند. (Floridi, 2006, p. 6) فلوریدی بر این باور است که همه موجودات، به دلیل برخورداری از بُعد و شأن اطلاعاتی، دارای ارزش ذاتی‌اند و بر همین اساس، برابری هستی‌شناختی دارند. (Floridi, 2013, p. 68) نابودی یا آسیب به هر موجودیت به معنای ایجاد نقص در اطلاعات است و هر عاملی که هستی را به نقصان یا عدم تبدیل کند، نوعی آنتروپی^{۱۱} یا اختلال محسوب می‌شود که در سپهر اطلاعات باید از آن جلوگیری شود. (Floridi, 2013, p. 65)

دیدگاه اخلاقی فلوریدی نیز بر اساس همین مبانی اطلاعاتی و هستی‌شناسانه شکل گرفته است. او یک نظریه اخلاقی هستی‌محور^{۱۲} ارائه می‌دهد که در آن، هستنده‌های غیرانسانی اعم از جانداران آگاه و غیرآگاه و غیرجانداران مانند ربات‌ها، سامانه‌های هوش مصنوعی و اطلاعاتی نیز موضوع اخلاق محسوب می‌شوند و دارای ارزش ذاتی‌اند. بدین ترتیب، عرصه اخلاق و کنش اخلاقی برای فلوریدی، عرصه‌ای جهان‌شمول است که تمام موجودات اطلاعاتی را شامل می‌شود. به عبارت دیگر، به تمامی موجودات اطلاعاتی که در فضا و سپهر اطلاعاتی قرار می‌گیرند، ارزش اخلاقی نسبت داده می‌شود و بنابراین آنها شایسته توجه و احترام اخلاقی خواهند بود. (Floridi, 2008, p. 15)

نظریه «اخلاق بدون ذهن» و ارائه مفهوم «سطح انتزاع»

⁶ Ontocentric

⁷ Infocentric

⁸ Informational Entity

⁹ clusters of information

¹⁰ Infosphere

¹¹ Entropy

¹² Ontocentric Ethics

فلوریدی با رویکرد هستی‌محور خود، تعاریف جدیدی از مفهوم «عامل» یا کنشگر اخلاقی^{۱۳} ارائه می‌دهد و پیشنهاد می‌کند که با ورود و مشارکت عامل‌های مصنوعی به گفتمان اخلاقی، باید مفهوم عامل اخلاقی را نیز گسترش داد تا همه هستنده‌های ساکن در سپهر اطلاعاتی از منظر غیرانسان‌محور تحلیل شوند.

(Floridi, 2013, p. 158)

او معتقد است که اخلاق استاندارد با نگاه محدود و انسان‌محور، تحت فشار تعامل انسان و کامپیوتر قرار دارد، زیرا در آن تنها انسان‌ها به عنوان عامل‌های اخلاقی شناخته می‌شوند (Floridi, 1999, p. 41) به همین دلیل، فلوریدی مستقیماً نظریه اخلاق استاندارد را مورد انتقاد قرار می‌دهد و پیشنهاد می‌کند که قرائت استاندارد عاملیت اخلاقی که بر ویژگی‌های ذهنی انسان، مانند آگاهی، قصدیت، اراده آزاد تمرکز دارد، کنار گذاشته شود و به جای آن، رویکردی جدید تحت عنوان «اخلاق بدون ذهن» ارائه شود.

(Floridi, 2013, p. 135; Floridi & Sanders, 2004, p. 351)

همانطور که خود او اشاره می‌کند مزیت بزرگ ایده او افزون بر اینکه به مشارکت و گسترش طبقه عامل‌های اخلاقی به عامل‌های اخلاقی مصنوعی می‌انجامد، همچنین به اندازه کافی پرسش‌های اخلاقی در مورد تعامل میان انسان و هوش مصنوعی را روشن می‌کند و درک بهتری از گفتمان اخلاقی و همچنین درک طیف وسیعی از مسائل اخلاقی جدید را در زمینه‌های غیرانسانی به دست می‌دهد. (Floridi & Sanders, 2004, p. 376)

در نظریه «اخلاق بدون ذهن» مفهوم محوری برای تعیین عاملیت اخلاقی، مفهوم و چارچوب توصیفی «سطح انتزاع»^{۱۴} است. به اعتقاد فلوریدی پدیده‌ها و واقعیات دارای سطوح گوناگونی هستند و می‌توان یک پدیده یا واقعیت را در سطوح گوناگونی مورد بررسی و تفسیر قرار داد. در سطوح گوناگون انتزاع، مشاهدات متفاوتی مرتبط هستند که باعث تغییر نگاه ما به جهان می‌شود. بر این اساس فلوریدی معتقد است که هر پدیده یا واقعیت دارای سطوح مختلفی از انتزاع است و برداشت‌های متفاوت از آن پدیده بسته به سطح انتخابی، تغییر می‌کند. (Floridi, 2013, p. 29) برای مثال، نهنگ ممکن است در یک سطح به عنوان پستاندار شناخته شود و در سطح دیگر به عنوان ماهی یا فردی گوجه‌فرنگی را جزء سیفی‌جات بداند و دیگری جزء دسته توت‌ها که این بستگی دارد به سطح انتزاعی که به وسیله آن، فرد موجودیت‌ها را تعریف می‌کند؛ مثلاً درباره گوجه‌فرنگی، تعریف و دسته‌بندی اول در سطح انتزاع یک کشاورز انجام می‌گیرد، و تعریف و دسته‌بندی دوم در سطح انتزاع یک گیاه‌شناس است. (Floridi, 2013, p. 139)

¹³ Moral Agent

¹⁴ Level of Abstraction (LoA)

فلوریدی معتقد است که تعیین عاملیت اخلاقی نیز بستگی به سطح انتزاع انتخاب شده دارد. در این دیدگاه، عاملیت اخلاقی دیگر نباید بر اساس ویژگی‌های انسانی مانند آگاهی، قصدیت یا اراده آزاد و غیره تعریف شود، بلکه باید از سطح انتزاع بالاتری بررسی شود. (Floridi & Sanders, 2004, p. 365) او نشان می‌دهد که سطح انتزاع مورد استفاده در رویکرد استاندارد بسیار پایین است و معیارها را نزدیک به انسان بالغ نگه می‌دارد، اما با افزایش سطح انتزاع، دیدگاه‌های کمتر انسان‌محور درباره عاملیت در دسترس قرار می‌گیرند. در این چارچوب، عامل‌های هوشمند مصنوعی می‌توانند در یک سطح انتزاع دیگر با معیارهای ساده‌تر به عنوان عامل‌های اخلاقی شناخته شوند. معیارهایی که فلوریدی پیشنهاد می‌کند عبارتند از: تعامل‌پذیری^{۱۵}، خودمختاری^{۱۶} و سازگاری^{۱۷} (Floridi, 2013, p. 140)

او تعامل‌پذیری را به معنای توانایی درک مستقیم و پاسخ به محیط یا به تعبیر دیگر تأثیرگذاری و تأثیرپذیری از محیط می‌داند. خودمختاری نیز به این معناست که یک عامل بتواند حالات خود را مستقل از محرک‌های محیطی تغییر دهد. سازگاری نیز اینگونه است که عامل شیوه‌های عمل خود را بر اساس تجارب گذشته تطبیق دهد. مفهوم عامل اخلاقی با توجه به سطح انتزاع مناسب می‌تواند بر روی هر عامل هوشمندی که به اندازه کافی خودمختار، سازگار و تعاملی باشد اطلاق شود و به عنوان عامل اخلاقی در نظر گرفته شود.

(Floridi, 2013, pp. 140-141)

بر اساس دیدگاه فلوریدی، موجودیتی مانند یک سنگ هیچ‌یک از این معیارها (تعامل‌پذیری، خودمختاری، سازگاری) را برآورده نمی‌کند. یک سنگ قادر به تعامل با محیط خود نخواهد بود. سنگ همچنین به طور واضح خودمختار نیست، زیرا نمی‌تواند کاری انجام دهد مگر اینکه عملی بر روی آن انجام شود، و بنابراین سازگار نیز نیست. اما موجودیتی مانند یک پاندول تنها خودمختاری را نشان می‌دهد و موجودیتی مانند یک ترموستات هم تعامل‌پذیر و هم سازگار است اما خودمختار نیست. از سوی دیگر، یک برنامه کامپیوتری مربوط به استخدام افراد که رزومه‌ها را بررسی می‌کند به نظر می‌رسد که تعاملی، خودمختار و سازگار است و می‌تواند به عنوان یک عامل اخلاقی در نظر گرفته می‌شود.

بر این اساس آنچه به عنوان عامل اخلاقی در نظر گرفته می‌شود دیگر نباید بر اساس مفاهیم و ویژگی‌های قرائت استاندارد عاملیت مورد بررسی قرار گیرد بلکه باید از یک منظر بالاتر ملاحظه شود. با معیارهای جدید هم انسان‌ها و هم عامل‌های مصنوعی می‌توانند عامل اخلاقی باشند یعنی هر دو به صورت تعاملی و بر اساس اطلاعات موجود در پاسخ به موقعیت جدیدی که با آن روبرو می‌شوند، عمل می‌کنند. هر دوی آنها در طول عمل، چندین بار رفتار

¹⁵ Interactivity

¹⁶ Autonomy

¹⁷ Adaptability

خود را بر اساس اطلاعات جدید موجود تغییر می‌دهند و در نتیجه به صورت خودمختار عمل می‌کنند و هر دو به صورت سازگار عمل می‌کنند یعنی آنها صرفاً از دستورات یا دستورالعمل‌های از پیش تعیین شده پیروی نمی‌کنند (Floridi, 2013, p. 148).

پاسخ فلوریدی به انتقادات وارد شده به نظریه «اخلاق بدون ذهن»

فلوریدی انتقاداتی را که به نظریه «اخلاق بدون ذهن» وارد شده است، در چند دسته طبقه‌بندی می‌کند و به هر یک پاسخ می‌دهد. در ادامه، مهم‌ترین این انتقادات همراه با پاسخ او به اختصار بیان می‌شود.

نقد غایت‌شناختی: محور اصلی این نقد آن است که در نظریه فلوریدی، عامل مصنوعی فاقد غایت یا هدف در معنای واقعی آن است. فلوریدی در پاسخ بیان می‌کند که این نقد قابل رفع است؛ زیرا عامل‌های مصنوعی می‌توانند رفتارهای هدفمند از خود نشان دهند و آن‌ها را بهبود بخشند. از این رو، ویژگی‌های غایت‌شناختی نمی‌توانند مبنای تمایز قاطع میان عامل انسانی و عامل مصنوعی باشند (Floridi, 2013, p. 148).

نقد مربوط به قصدیت: بر اساس این نقد، عامل مصنوعی فاقد حالات آگاهانه قصدی است. فلوریدی در پاسخ استدلال می‌کند که حالات قصدی شرطی مطلوب، اما غیرضروری برای تحقق عاملیت اخلاقی هستند. به نظر او، رویکرد استاندارد در اخلاق، وجود نوعی دسترسی ویژه به حالات ذهنی یا قصدی درونی را در انسان مفروض می‌گیرد، در حالی که چنین دسترسی‌ای به جز شاید حالات ذهنی خودمان در عمل به‌سادگی قابل تضمین یا احراز نیست (Floridi, 2013, p. 149).

نقد مربوط به اراده آزاد: این نقد مدعی است که عامل مصنوعی فاقد اراده آزاد است. پاسخ فلوریدی به این انتقاد مشابه پاسخ او به نقد پیشین است. او معتقد است که هرگونه نقد مبتنی بر دسترسی به حالات ذهنی خاص، با محدودیت‌های معرفتی مواجه است. از آنجا که دسترسی مستقیم و قابل اتکایی به اراده آزاد وجود ندارد، نمی‌توان آن را به‌عنوان شرط تعیین‌کننده عاملیت اخلاقی در نظر گرفت (Floridi, 2013, p. 149).

با ملاحظه مواردی که در بالا بیان شد، می‌توان دریافت که با وجود نوآوری‌های نظری دیدگاه فلوریدی در گسترش قلمرو اخلاق و توجه به نقش روزافزون عامل‌های هوشمند مصنوعی در سپهر اطلاعات، نظریه «اخلاق بدون ذهن» همچنان با چالش‌ها و ابهامات بنیادین فلسفی مواجه است. مهم‌ترین مسئله آن است که آیا می‌توان بدون در نظر گرفتن ویژگی‌های ذهنی همچون قصدیت و اراده آزاد، همچنان از «عاملیت اخلاقی» به معنای دقیق سخن گفت یا خیر. به نظر می‌رسد فلوریدی با بسط گسترده مفهوم عاملیت اخلاقی و بازتعریف معیارهای آن، مرز میان عاملیت

اخلاقی و صرفاً یک سامانه کارکردی را تا حدی مخدوش می‌کند. در ادامه، این نظریه در قالب سه نقد بنیادین مورد تحلیل و ارزیابی انتقادی قرار می‌گیرد و نشان داده می‌شود که چارچوب نظریه «اخلاق بدون ذهن» و پاسخ‌های فلوریدی به انتقادات واردشده، در مقایسه با روایت استاندارد عاملیت اخلاقی، از ظرفیت تبیینی قانع‌کننده‌تری برخوردار نیست. در نهایت، تلاشی برای بازسازی دفاع فلسفی از نظریه «اخلاق بدون ذهن» صورت می‌گیرد و این بازسازی نیز به صورت انتقادی مورد ارزیابی قرار خواهد گرفت.

1. تقلیل‌گرایی کارکردی در تبیین عاملیت اخلاقی

فلوریدی در پاسخ به «نقد غایت‌شناختی» بروز رفتارهای هدفمند از عامل‌های مصنوعی را برای تحقق عاملیت اخلاقی در آنها کافی می‌داند. این دیدگاه متأثر از بدنه اصلی نظریه «اخلاق بدون ذهن» است که او در تلاش برای گسترش قلمرو عاملیت اخلاقی و عبور از روایت انسان‌محور سنتی، معیارهایی همچون تعامل‌پذیری، خودمختاری و سازگاری را به عنوان شروط اصلی عاملیت اخلاقی معرفی می‌کند. بر اساس معیارهای مورد نظر فلوریدی، هر سامانه‌ای که بتواند با محیط تعامل کند، مستقل عمل نماید و رفتار خود را با شرایط جدید سازگار سازد، می‌تواند در زمره عامل‌های اخلاقی قرار گیرد. در این چارچوب، برخورداری از حالات ذهنی، ضروری تلقی نمی‌شوند. با این حال، پرسش اساسی آن است که آیا این نظریه فراتر از توصیفی صرفاً کارکردی از عاملیت اخلاقی ارائه می‌دهد؟ به نظر می‌رسد نظریه فلوریدی در پاسخ به این پرسش با دشواری جدی مواجه است. بسیاری از اشیاء، نهادها و سامانه‌های فناورانه می‌توانند داده‌های محیطی را دریافت کنند، نسبت به آنها واکنش نشان دهند و رفتار خود را با شرایط جدید تطبیق دهند و منشأ پیامدهای اخلاقی باشند، اما آنها فاقد «عمل» به معنای دقیق کلمه‌اند و آنچه از آنها صادر می‌شود صرفاً رفتاری ناشی از سازوکارهای علی و فیزیکی است. برای مثال، ممکن است یک سامانه هوش مصنوعی در فرآیند استخدام یا تشخیص پزشکی تصمیم‌هایی اثرگذار اتخاذ کند و حتی پیامدهای اخلاقی مهمی نیز بر تصمیمات آن مترتب باشد، اما این امر نشان نمی‌دهد که سامانه مذکور مفاهیمی چون عدالت، مسئولیت، آسیب یا خیر اخلاقی را فهم می‌کند. چنین سامانه‌ای صرفاً اطلاعات را پردازش می‌کند و بر اساس الگوریتم‌های از پیش تعیین‌شده واکنش نشان می‌دهد.

به اعتقاد ما این معیارها به گونه‌ای صورت‌بندی شده‌اند که سبب شده‌اند تمایز ماهوی میان «عمل اخلاقی» و «رفتار اخلاقی» به درستی تبیین نشود و در نتیجه اگرچه این معیارها می‌توانند دامنه عاملیت اخلاقی را گسترش دهند، اما هم‌زمان محتوای هنجاری و اخلاقی مفهوم عاملیت اخلاقی را نیز تا حد زیادی رقیق و تهی می‌کنند. در

واقع، معیارهای پیشنهادی فلوریدی بیش از آنکه ناظر به عاملیت اخلاقی در معنای دقیق فلسفی و در سطح متافیزیکی و ذهنی باشند به سطح کارکردی و حداقلی تقلیل می‌یابند.

در مقابل، روایت استاندارد عاملیت اخلاقی صرف صدور رفتار یا تأثیرگذاری علی بر محیط را برای انتساب عاملیت اخلاقی کافی نمی‌داند، بلکه عاملیت را مستلزم توانایی انجام «عمل» اخلاقی به معنای دقیق کلمه می‌داند. بر این اساس، عامل اخلاقی تنها موجودی واکنش‌گر یا پردازشگر اطلاعات و در عین حال هدفمند نیست، بلکه موجودی است که می‌تواند عمل یا کنش خود را در افق مفاهیمی چون خیر، وظیفه، عدالت و مسئولیت فهم و ارزیابی کند که این امر مستلزم برخورداری از حالات ذهنی است. بر این اساس، مهم‌ترین پیش‌فرض بنیادین در نظریه استاندارد عاملیت اخلاقی آن است که میان عاملیت و ذهنیت رابطه‌ای ضروری برقرار است و از این منظر، تقلیل عاملیت به صرف رفتار ناکافی است و نسبت دادن عاملیت به موجودی فاقد حالات ذهنی صرفاً استعاری خواهد بود. به تعبیری می‌توان به صورت خلاصه گفت: موجودی عامل است اگر و تنها اگر بتواند حالات ذهنی‌ای داشته باشد که منشأ انجام عمل باشند.

از همین رو، بسیاری از فیلسوفان مدافع روایت استاندارد میان «رفتار» و «عمل» تمایزی بنیادین قائل شده‌اند و معتقدند که هر رفتار حتی پیچیده‌ای را نمی‌توان با عمل اخلاقی یکسان دانست. در همین چارچوب، هیمما و همچنین یاکوب تصریح می‌کنند که بنیاد عاملیت، ظرفیت ایجاد عمل و کنش است و یک موجود تنها زمانی می‌تواند عامل تلقی شود که قادر به انجام عمل باشد، نه صرفاً بروز رفتار. به اعتقاد آنها، برای اینکه یک رویداد «عمل» محسوب شود، علت آن باید حالات درونی عامل باشد و این حالات درونی نیز، بنا بر فرض رایج در سنت فلسفی، حالات ذهنی‌اند؛ این حالات ذهنی شامل طیفی از باورها، خواسته‌ها، انگیزه‌ها، تمایلات، قضاوت‌ها و نیت‌ها هستند که همگی ذیل مفهوم قصدیت قرار می‌گیرند و نمونه‌هایی از آن هستند. (Himma, 2009, pp. 19-20; Jacob, 2019, pp. 1-5)

برای مثال، توضیح اینکه چرا شخصی یک لیوان شیر می‌نوشد، معمولاً با ارجاع به حالات قصدی او صورت می‌گیرد؛ مانند باور او به رفع تشنگی یا نگرانی‌اش از کم‌آبی بدن. بر همین مبنا، بسیاری از متفکران طرفدار دیدگاه استاندارد بر نقش قصدیت به طور کلی تأکید کرده‌اند. شلوسر تصریح می‌کند که عمل مستلزم قصدیت است و قصدیت نیز به علّیتی وابسته است که از حالات ذهنی عامل ناشی می‌شود. (Schlosser, 2015, p. 1)

جانسون نیز معتقد است که برای اینکه موجودی بتواند «عمل» کند و نه صرفاً «رفتار» از خود نشان دهد، باید دارای حالات ذهنی درونی، به‌ویژه قصد عمل، باشد (Johnson, 2006, pp. 198-200). از نظر او، رفتار قصدی (عمل) نوعی رابطه علی پیچیده دارد که آن را از رفتار غیرقصدی (رفتار صرف) متمایز می‌کند؛ زیرا رفتار قصدی

(عمل) رفتاری است که از حالات ذهنی درونی سرچشمه می‌گیرد و از همین رو، افزون بر تبیین علّی، قابلیت تبیین عقلانی نیز دارد. به بیان دیگر، همه رفتارها را می‌توان با ارجاع به علل توضیح داد، اما تنها اعمال را می‌توان با ارجاع به باورها، خواسته‌ها، انگیزه‌ها و سایر حالات ارادی عامل تبیین کرد.

بری نیز استدلال می‌کند که تعامل اخلاقی واقعی تنها به وسیله عمل قصدی صورت می‌پذیرد و معتقد است که از آنجا که عمل با ارجاع به قصد توضیح داده می‌شود ما از خلال آن به انگیزه‌ها، توجیهات و ساختار ذهنی عامل دست می‌یابیم و همین امر مبنای ارزیابی‌های اخلاقی ماست. پاسخ‌های ما به یک عمل معمولاً با پاسخ به رویدادی غیر از آن متفاوت است یعنی ما در ارزیابی اعمال نه تنها به پیامدهای فیزیکی آنها بلکه به باورها، خواسته‌ها و سایر حالات ذهنی زیربنای آنها نیز توجه اخلاقی و اجتماعی داریم. (Brey, 2013, p. 132)

میسلهورن اما نکته مهم دیگری را در ارتباط با عمل و حالات درونی بیان می‌کند و آن این است که میان «عمل» با مفهوم «خودآغازگری» پیوندی مستقیم وجود دارد. عامل تنها زمانی خودآغازگر تلقی می‌شود که منشأ عمل او در درون خود او باشد، نه در عوامل و محرک‌های بیرونی. به تعبیر میسلهورن، آغازکنندگان عمل باید ساختار درونی و تمایلات عامل باشند، نه صرفاً رویدادهای خارجی. (Misselhorn, 2018, pp. 163-164) بنابراین، منشأ عمل اخلاقی نه صرف واکنش به داده‌های محیطی، بلکه نوعی حالت ذهنی و درونی است.

2. حذف نقش بنیادین حالات ذهنی

همان‌گونه که ملاحظه شد، تلاش فلوریدی در نظریه «اخلاق بدون ذهن» مبتنی بر کنار گذاشتن نقش ویژگی‌هایی همچون آگاهی، قصدیت و حالات ذهنی در تبیین عاملیت اخلاقی است. فلوریدی با نقد روایت استاندارد عاملیت اخلاقی، این ویژگی‌ها را شروطی غیرضروری برای عاملیت می‌داند و معتقد است که تمرکز بر آنها اخلاق را به‌طور ناموجهی انسان‌محور کرده است. او در پاسخ به «نقد مربوط به قصدیت» و «نقد مربوط به اراده آزاد»، استدلال می‌کند که معمول‌ترین معیارهای بررسی عاملیت اخلاقی در دیدگاه استاندارد مبتنی بر ویژگی‌های ذهنی است که قابل دسترس نیستند. از آنجا که دسترسی مستقیم به این حالات ذهنی برای ما ممکن نیست، نمی‌توان با قطعیت ثابت کرد که آیا هیچ موجودی در واقع جنبه‌های درونی ذهنیت را که عموماً برای جایگاه اخلاقی در دیدگاه استاندارد ضروری تلقی می‌شوند، داراست یا خیر.

به نظر ما، این استدلال با اشکالات جدی مواجه است. دشواری معرفت‌شناختی در احراز یک ویژگی، دلیلی بر نفی و بی‌اعتباری هستی‌شناختی آن ویژگی نیست. برای مثال، انسان‌ها هرگز درد، رنج یا حالات ذهنی دیگران را به‌طور مستقیم تجربه نمی‌کنند، اما این حالات درونی نقشی بنیادین در قضاوت‌ها و واکنش‌های اخلاقی دارند. به همین

ترتیب، عدم دسترسی مستقیم به قصدیت، آگاهی یا اراده آزاد نمی‌تواند دلیلی کافی برای حذف آنها از مفهوم عاملیت اخلاقی باشد. افزون بر این، بسیاری از مفاهیم اخلاقی همچون مسئولیت، تعهد، نیت خیر یا سوءنیت، اساساً بر وضعیت‌های ذهنی مبتنی هستند که در واقع، بخش بزرگی از روابط اخلاقی و حقوقی دقیقاً بر فرض وجود همین حالات ذهنی بنا شده است. در نتیجه، نادیده گرفتن، حذف یا تقلیل نقش حالات و عناصر ذهنی از عاملیت اخلاقی، خطر تهی شدن ساحت اخلاق از محتوای هنجاری آن را به همراه دارد و می‌تواند پیامدهای فلسفی مهمی در تبیین مسئولیت اخلاقی و حقوقی ایجاد کند.

3. وابسته ساختن عاملیت اخلاقی به مفهوم سطح انتزاع

همانگونه که ملاحظه شد یکی از مهم‌ترین ابزارهای مفهومی فلوریدی برای دفاع از عاملیت اخلاقی عامل‌های هوشمند مصنوعی، مفهوم «سطح انتزاع» است. به باور او، با افزایش سطح انتزاع و فاصله گرفتن از معیارهای انسان‌محور، امکان انتساب عاملیت اخلاقی به سامانه‌های هوشمند نیز فراهم می‌شود. با این حال، به کارگیری این معیار نیز با اشکالاتی مواجه است.

یکی از مهم‌ترین اشکالاتی که به نظریه فلوریدی وارد شده، نقدی است که جانسون و میلر مطرح کرده‌اند. به اعتقاد آنان، فلوریدی با پذیرش این ایده که یک موجودیت ممکن است در یک سطح انتزاع عامل اخلاقی محسوب شود و در سطحی دیگر چنین نباشد، مفهوم عاملیت اخلاقی را به مفهومی نسبی و وابسته به چارچوب توصیف تقلیل می‌دهد. از نظر آنان، اخلاق خود ناظر به سطح خاصی از انتزاع با مفاهیم و معیارهای معین است؛ بنابراین نمی‌توان بدون ارائه معیاری مستقل، سطوح مختلف انتزاع را به دلخواه مبنای داوری اخلاقی قرار داد. در نتیجه، این ادعا که می‌توان گستره مفاهیم اخلاقی را صرفاً با تغییر سطح انتزاع کاهش یا افزایش داد، نوعی مصادره به مطلوب به شمار می‌آید. (Johnson & Miller, 2008, p. 128) بر این اساس، نظریه فلوریدی توضیح روشنی درباره اینکه چرا باید یک سطح انتزاع خاص از میان سطوح متعدد ممکن برگزیده شود ارائه نمی‌کند. فلوریدی ظاهراً فرض می‌کند که عرصه اخلاق و معیارهای اخلاقی دارای سطوح گوناگون انتزاع‌اند، نه اینکه ذاتاً به یک سطح مشخص تعلق داشته باشند. اما اگر انتخاب سطح انتزاع تا حد زیادی وابسته به اهداف توصیفی ناظر باشد، در آن صورت عاملیت اخلاقی نیز به امری نسبی و وابسته به چارچوب توصیف تبدیل می‌شود.

اشکال دوم نیز به اندازه اشکال نخست جدی است. می‌توان تبیین فلوریدی از معیار «سطح انتزاع» را نوعی لغزش مفهومی دانست. او از این گزاره که :

« X در یک سطح انتزاع، تعاملی، خودمختار و سازگار است و در نتیجه عامل اخلاقی محسوب می‌شود»

به این نتیجه می‌رسد که: « X واقعاً یک عامل اخلاقی است».

به بیان دیگر، فلوریدی از این که چیزی در یک سطح توصیف، واجد یک ویژگی باشد، به این نتیجه گذار می‌کند که آن چیز به نحو کلی و واقعی واجد آن ویژگی است. لازمه منطقی تحلیل او این است که سطح انتزاع بتواند ماهیت یک موجود را تغییر دهد. در حالی که به نظر نمی‌رسد تغییر در سطح توصیف یک موجود، لزوماً ماهیت وجودی آن موجود را دگرگون سازد. به بیان دیگر، این که یک سامانه هوش مصنوعی در سطحی انتزاعی تر، تعاملی، خودمختار و سازگار توصیف شود، به خودی خود موجب پیدایش ویژگی‌های مهمی همچون آگاهی، فهم اخلاقی یا قصدیت واقعی در آن سامانه نخواهد شد. از این منظر، می‌توان معیار سطح انتزاع در دیدگاه فلوریدی را بیشتر ابزاری معرفت‌شناختی برای توصیف و سامان‌دهی پدیده‌ها دانست، نه معیاری هستی‌شناختی برای تعیین نوع وجود اشیا و در نتیجه اثبات عاملیت اخلاقی آنها. برای مثال، توصیف یک خودروی خودران در یک سطح انتزاع ممکن است آن را یک سامانه تصمیم‌گیر نشان دهد، اما چنین توصیفی ماهیت محاسباتی آن را به ماهیت یک موجود آگاه و دارای ذهن تبدیل نمی‌کند.

بنابراین، همان‌گونه که ملاحظه می‌شود، فلوریدی با به کارگیری این مفهوم نمی‌تواند تبیینی اقناع‌کننده و برخوردار از مزیتی آشکار نسبت به دیدگاه استاندارد ارائه دهد. دیدگاه استاندارد در این زمینه، مبنایی عینی‌تر، هستی‌شناختی‌تر و کم‌تر نسبی‌گرایانه برای عاملیت اخلاقی فراهم می‌آورد که با شهودهای متعارف اخلاقی نیز سازگارتر به نظر می‌رسد.

بازسازی دفاع فلسفی از نظریه «اخلاق بدون ذهن» و ارزیابی انتقادی آن

در برابر انتقاداتی که به نظریه «اخلاق بدون ذهن» فلوریدی وارد شده است، ممکن است مدافعان این دیدگاه استدلال کنند که بخش قابل توجهی از این نقدها همچنان در چارچوب پیش‌فرض‌های محافظه‌کارانه و انسان‌محور روایت استاندارد از عاملیت اخلاقی باقی مانده‌اند و دقیقاً به همین دلیل، توان درک دگرگونی‌های مفهومی ناشی از ظهور عامل‌های هوشمند مصنوعی را ندارند. از این منظر، اصرار بر مؤلفه‌های آگاهانه و ذهنی مبتنی بر این پیش‌فرض سنتی است که تنها موجوداتی با ساختار روان‌شناختی مشابه انسان می‌توانند در قلمرو اخلاق قرار گیرند؛ حال آنکه فلوریدی اساساً در پی نقد این انحصارگرایی انسان‌محور و گسستن پیوند سنتی میان ذهنیت و عاملیت اخلاقی است. بر این اساس، نظریه «اخلاق بدون ذهن» را می‌توان تلاشی برای بازتعریف مفهوم عاملیت

اخلاقی در جهانی دانست که در آن سامانه‌های هوشمند به‌طور فزاینده‌ای در فرایندهای تصمیم‌گیری، تعاملات اجتماعی و تولید پیامدهای اخلاقی مشارکت دارند.

با این حال، چنین دفاعی نیز نمی‌تواند به‌طور کامل دشواری‌های بنیادین این نظریه را برطرف سازد. هرچند فلوریدی روایت استاندارد عاملیت اخلاقی را به انسان‌محوری متهم می‌کند، اما پیوند میان عاملیت اخلاقی و مؤلفه‌های ذهنی را نمی‌توان صرفاً یک پیش‌داوری انسان‌شناختی دانست. تأکید روایت استاندارد بر ویژگی‌هایی همچون قصدیت، آگاهی و سایر مؤلفه‌های ذهنی، مبتنی بر تحلیل مفهومی خود «عمل اخلاقی» است. همان‌گونه که ملاحظه شد، تمایز بنیادینی میان «رفتار» و «عمل» وجود دارد. آنچه یک رویداد را به «عمل» تبدیل می‌کند صرفاً وقوع یک فرایند علی یا بروز رفتاری هدفمند نیست، بلکه امکان تبیین آن بر اساس باورها، انگیزه‌ها، خواسته‌ها و مقاصد عامل است؛ به این معنا که منشأ عمل باید در ساختار ذهنی و قصدی خود عامل باشد (نک: ذاکری، 1394، فصل 1؛ Davidson, 1980, pp. 43-62; Hornsby, 1980, Ch. 1; O'Connor, 2000, p. 86; Ginet, 1990, pp. 11-30)، نه صرفاً در سازوکارهای رفتاری یا پردازش اطلاعات. بر این اساس، حذف مؤلفه‌های ذهنی از مفهوم عاملیت اخلاقی صرفاً به گسترش دامنه اخلاق نمی‌انجامد، بلکه خطر تهی‌سازی این مفهوم از محتوای هنجاری و فلسفی آن را در پی دارد. بنابراین مسئله صرفاً توسعه دامنه عاملیت نیست، بلکه حفظ شرایط عاملیت اهمیت دارد؛ یعنی شرایطی که در آن نسبت‌دادن عاملیت اخلاقی همچنان معنا، انسجام و کارکرد هنجاری خود را حفظ کند. اگر هر سامانه‌ای که قادر به دریافت و پردازش داده‌های محیطی و واکنش بر اساس آن باشد «عامل اخلاقی» تلقی شود، مفهوم عاملیت به سطحی از کارکردگرایی تقلیل می‌یابد که قدرت تبیینی خود را از دست می‌دهد. در عین حال، این نقد را نباید به‌عنوان نوعی محافظه‌کاری ساده‌انگارانه یا مقاومت در برابر تحولات فناورانه تلقی کرد. مقصود از دفاع از روایت استاندارد، تأکید بر ضرورت حفظ تمایز میان «عاملیت اخلاقی» و «نقش اخلاقی» است. بدون تردید، عامل‌های هوشمند مصنوعی نقش‌های اخلاقی مهمی ایفا می‌کنند و بخشی از بافت اخلاقی جامعه را شکل می‌دهند. این سامانه‌ها صرفاً ابزارهای خنثی نیستند، بلکه ساختارهایی‌اند که می‌توانند رفتار انسانی را هدایت، تقویت یا محدود کنند و در تولید پیامدهای اخلاقی نیز نقش داشته باشند. برای مثال، سامانه‌ای که راننده را به بستن کمربند ایمنی ترغیب می‌کند یا الگوریتمی که تبعیض مبتنی بر نژاد یا مذهب را در فرآیند اعطای وام حذف می‌کند، در جهت بهبود پیامدهای اخلاقی عمل می‌کنند. با این حال، مسئله اصلی زمانی پدید می‌آید که از این «نقش اخلاقی» نتیجه گرفته شود که این سامانه‌ها دارای «عاملیت اخلاقی» هستند. از این منظر، روایت استاندارد بر این نکته تأکید دارد که میان عاملیت اخلاقی انسان و نقش‌های اخلاقی موجودات غیرانسانی تفاوتی کیفی وجود دارد. این تمایز امکان می‌دهد هم اهمیت اخلاقی فناوری به رسمیت شناخته شود و هم ویژگی‌های متمایزکننده عاملیت انسانی همچون آگاهی، قصدیت، اراده آزاد حفظ گردد. این نکته زمانی اهمیت بیشتری می‌یابد که توجه کنیم که در روایت استاندارد عاملیت اخلاقی، انتساب عاملیت اخلاقی به انسان به عنوان نمونه

اعلای این مفهوم نیز با قیود و تمایزات متعدد همراه است. در این چارچوب، تنها انسان‌های بالغ و برخوردار از قوای عقلانی به‌عنوان عامل اخلاقی در نظر گرفته می‌شوند؛ زیرا آنها توان استدلال اخلاقی و تمایز میان درستی و نادرستی را دارند. در مقابل، انسان‌هایی که به دلیل اختلالات شناختی یا آسیب‌های شدید ذهنی از این ظرفیت‌ها محروم‌اند، معمولاً در زمره عامل‌های اخلاقی قرار نمی‌گیرند. برای نمونه می‌توان به مبتلایان به آلزایمر در مراحل متوسط تا پیشرفته یا افراد دارای اختلالات شدید روانی اشاره کرد که توانایی‌های لازم برای ارزیابی و داوری اخلاقی را تا حد زیادی از دست داده‌اند (Watson, 2013, pp. 9–10). به‌طور مشابه، کودکان خردسال و نوزادان نیز به دلیل فقدان ظرفیت‌های شناختی و عقلانی کافی، واجد عاملیت اخلاقی به معنای دقیق آن محسوب نمی‌شوند. افزون بر این، حتی درباره انسان‌های بالغ و سالم نیز انتساب مسئولیت اخلاقی منوط به آن است که کنش‌های آنان آگاهانه و به‌صورت آزادانه و ارادی انجام شده باشد (Onyeukaziri, 2023, p. 12).

نتیجه‌گیری

نظریه «اخلاق بدون ذهن» فلوریدی را می‌توان یکی از مهم‌ترین تلاش‌های معاصر برای بازاندیشی در مفهوم عاملیت اخلاقی در عصر فناوری‌های هوشمند دانست. اهمیت این نظریه در آن است که می‌کوشد با عبور از چارچوب انسان‌محور سنتی، نسبت انسان، فناوری و اخلاق را در سپهر اطلاعاتی جدید بازتعریف کند و جایگاهی برای عامل‌های مصنوعی در گفتمان اخلاقی قائل شود. با این حال، بررسی انتقادی این دیدگاه نشان می‌دهد که گسترش قلمرو عاملیت اخلاقی در نظریه فلوریدی، با نوعی تقلیل‌گرایی مفهومی همراه است که در نهایت مرز میان «عمل اخلاقی» و «رفتار اخلاقی» را مخدوش می‌سازد. نخست آنکه معیارهایی مانند تعامل‌پذیری، خودمختاری و سازگاری، هرچند می‌توانند برخی قابلیت‌های سامانه‌های هوشمند را توضیح دهند، اما برای تبیین عاملیت اخلاقی در معنای دقیق فلسفی کافی نیستند؛ زیرا این معیارها لزوماً مستلزم فهم هنجاری، ادراک اخلاقی یا قصدیت واقعی نیستند. در نتیجه، نظریه فلوریدی بیش از آنکه تبیینی از «عمل اخلاقی» ارائه دهد، نوعی توصیف کارکردی از رفتار سامانه‌های پیچیده عرضه می‌کند. دوم آنکه کنار گذاشتن نقش آگاهی، قصدیت، اراده آزاد و سایر حالات ذهنی از مفهوم عاملیت اخلاقی، به تضعیف بنیان‌های هنجاری اخلاق می‌انجامد. بسیاری از مفاهیم بنیادین اخلاقی و حقوقی، همچون مسئولیت، نیت، تعهد و سرزنش اخلاقی، بر فرض وجود حالات درونی و ذهنی استوارند و حذف این مؤلفه‌ها، مفهوم عاملیت اخلاقی را از محتوای هنجاری آن تهی می‌کند. سوم آنکه معیار «سطح انتزاع» نیز نمی‌تواند مبنایی قانع‌کننده برای اثبات عاملیت اخلاقی فراهم آورد؛ زیرا تغییر در سطح توصیف یک موجود، لزوماً موجب تغییر در ماهیت هستی‌شناختی آن نمی‌شود. از این منظر، سطح انتزاع را باید ابزاری معرفت‌شناختی برای توصیف پدیده‌ها دانست، نه معیاری هستی‌شناختی برای تعیین واقعیت عاملیت

اخلاقی. همچنین هرچند سامانه‌های هوش مصنوعی می‌توانند منشأ نقش و پیامدهای اخلاقی مهمی باشند اما این امر به خودی خود برای انتساب عاملیت اخلاقی کامل به آن‌ها کفایت نمی‌کند. از این رو، به نظر می‌رسد نظریه فلوریدی، علی‌رغم نوآوری‌های مفهومی و اهمیت آن در گشودن افق‌های جدید در اخلاق فناوری، هنوز نتوانسته است جایگزینی اقناع‌کننده برای روایت استاندارد عاملیت اخلاقی فراهم آورد.

مقاله پذیرش شده

ذاکری، م. (1394). درآمدی به فلسفه عمل. سمت.

References:

- Behdadi, D., & Munthe, C. (2018). Artificial moral agency: Philosophical assumptions, methodological challenges, and normative solutions. Manuscript submitted for publication.
- Brey, P. (2013). From moral agents to moral factors: The structural ethics approach. The moral status of technical artefacts, Springer: 125-142.
- Coeckelbergh, M. 2010. "Robot rights? Towards a social-relational justification of moral consideration." *Ethics and Information Technology* 12(3): 209–221. doi:10.1007/s10676-010-9235-5
- Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Clarendon Press.
- Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology*, 1(1) 37–56.
- Floridi, L. (2006). Information ethics, its nature and scope. In J. van den Hoven & J. Weckert (Eds.), *Moral philosophy and information technology* (pp. 40 - 65). Cambridge University Press.
- Floridi, L. (2008). Foundations of information ethics. In H. Tavani & K. Himma (Eds.), *Information and computer ethics* (pp. 3–23). John Wiley & Sons.
- Floridi, L. (2013). *The ethics of information*. Oxford University Press.
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14 349–379.
- Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.
- 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 information technology*, 11(1), 19-29.
- Hornsby, J. (1980). *Actions*. London: Routledge and Kegan Paul.
- Jacob, Pierre 2019, 'Intentionality' The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2019/entries/intentionality/>
- Johnson, D. G. (2006). "Computer systems: Moral entities but not moral agents." *Ethics and Information Technology* 8(4): 195-204.
- Johnson, D. G. and K. W. Miller (2008). "Un-making artificial moral agents." *Ethics and Information Technology* 10(2): 123-133.
- O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.
- McKenna M. (2012). *Conversation and Responsibility*. New York: Oxford University Press.
- Misselhorn, C. (2018). Artificial morality. Concepts, issues and challenges. *Society*, 55(2), 161-169.
- Onyeukaziri, J. N. (2023). "Action and agency in artificial intelligence: a philosophical critique."

- Schlosser M. (2015). Agency – Stanford encyclopedia of philosophy. Available at: <https://plato.stanford.edu/entries/agency>
- Watson, G. (2013). "Moral agency." International encyclopedia of ethics.

مقاله پذیرش شده