فصلنامه علمی ـ پژوهشی پژوهش‌نامه اخلاق
دوره چهاردهم، بهـار ۱۴۰۰، شماره ۵۱
صفحات: ۹۲ ـ ۶۹

PAZHUOHESH NAMEH-E AKHLAGH
A Research Journal
Vol. 14, Spring 2021, No. 51

چالش‌های فراروی کاربست اخلاق در ماشین‌های هوشمند؛ با تمرکز بر رویکرد اصل‌گرایی در اخلاق[۱]

نرگس کریمی واقف*
زهره عبدخدائی**

چکیده

با ظهور تکنولوژی هوش‌مصنوعی و حضور ماشین‌های هوشمند در بسیاری از ابعاد زندگی انسان، پرسش از امکان عملکرد اخلاقی ماشین‌های هوشمند و همچنین چگونگی و نحوه کاربست اخلاق در ماشین یکی از دغدغه‌های معاصر به شمار می‌آید. یکی از رویکردهای مهم در فلسفه اخلاق، اصل‌گرایی اخلاقی است که قابل بررسی با روش بالا به پایین بوده که یکی از رویکردهای متداول در پیاده‌سازی اصول در ماشین است. نوشتار حاضر در پی پاسخ به این پرسش است که آیا کاربست اخلاقی براساس اصل‌گرایی در اخلاق با روش پیاده‌سازی بالا به پایین امکان‌پذیر خواهد بود یا خیر؟ در این راستا، هدف بررسی و تحلیل بینش‌ها و چالش‌های پیش‌روی این رویکرد و روش، تحلیل و نقد ابعاد مساله و نظریات پیرامون آن است. بدین منظور پس از پاره‌ای تعاریف، دو رویکرد اخلاقیِ مطرح بیان گردیده، انطباق آن بر رویکرد بالا به پایین بررسی شده و سپس با استعانت از انتقادات و نظرات دیگر اندیشمندان، چالش‌ها و ناکارآمدی‌های این رویکرد به تفصیل شرح داده خواهد شد.

واژگان کلیدی

اصل‌گرایی اخلاقی، رویکرد بالا به پایین، اخلاق ماشین، اخلاق تکنولوژی، هوش مصنوعی.

---

۱. این پژوهش مورد حمایت مادی و معنوی بنیاد ملی نخبگان می‌باشد.
*. دانشجوی دکتری فلسفه دین، دانشگاه تهران. n.karimi.v@ut.ac.ir
**. دانش‌آموخته دکتری فلسفه تطبیقی دانشگاه علامه طباطبائی و مدرس دروس معارف اسلامی. zohreh.a@gmail.com
تاریخ دریافت: ۱۳۹۹/۱۲/۱۳ تاریخ پذیرش: ۱۴۰۰/۰۲/۲۹

طرح مسئله

ظهور اتومبیل‌های خودران، سلاح‌های جنگنده خودمختار و حتی دستیارهای هوشمند تلفن همراه، همگی حاکی از حضور گسترده و فعال سیستم‌های هوشمند در زندگی بشر است. سلاح‌هایی که به اندازه یک حشره بوده، تشخیص چهره داده و به فرد مورد نظر شلیک می‌کنند.[۱] شبکه‌های اجتماعی که براساس هر کلیلک کاربر، شناختی از او حاصل نموده، پروفایلی برای او بازکرده و علائق و سلایق او را شناخته و قادرند با پیشنهادات برمبنای علایق فرد، سلائق او را تغییر داده و حتی هدایت کنند[۲] و یا با استفاده از تصاویر حقیقی یک شخص خاص، تصاویر و فیلم‌های غیرواقعی جعل کنند به‌طوری‌که تشخیص آنها بسیار دشوار است.[۳]

دستیارهای هوشمند تلفن همراه که به خوبی انسان را شناخته، به روحیات او آشنایی داشته، به او پیشنهاد داده و برای او تصمیم‌سازی می‌نمایند. اگرچه این تصمیم‌سازی‌ها در بسیاری از مواقع مطلوب فرد است، اما آگاهی از تمامی اطلاعات فرد و تصمیم‌گیری برای او، نه تنها نقض حریم خصوصی افراد محسوب می‌شود، بلکه می‌تواند به نوع جدیدی از استبداد اجتماعی و یا حتی فرامنطقه‌ای تبدیل شود.

هرچه سطح خودمختاری در سیستم‌های هوشمند بالاتر رود، به تبع آن دغدغه عملکرد اخلاقی نیز پررنگ‌تر می‌شود. بنابراین اعتماد به هوش‌مصنوعی و سپردن مسئولیت‌های متعدد به آن، بدون در نظر گرفتن و رعایت اصول اخلاقی می‌تواند تهدید محسوب شود.[۴]

حال پرسش اینجاست که رعایت اصول اخلاقی توسط دستگاه هوشمند به چه معناست؟

---

۱. ایوب قرا، فعال سیاسی و وزیر ارتباطات رژیم اشغالگر قدس، در سال ۲۰۱۷ اعلام کرد که اسرائیل در حال گسترش تکنولوژی سلاح‌های هوشمند بوده و تاکنون توانسته نمونه‌هایی به کوچکی یک حشره را تولید کند.

'Development in Israel of terrorist - killing robots is no state secret', Udi Shaham, July 20, 2019, https: //www. jpost. com/Israel_News/Politics_And_Diplomacy/Kara_I_wasnt_revealing_state_secrets_about_the_robots_482616; and, Mini - nukes and mosquito - like robot weapons being primed for future warfare, Jeff Daniels, July 20, 2019, https: //www. cnbc. com/2017/03/17/mini_nukes_ and_inspect_bot_weapons_being_primed_for_future_warfare. html

۲. برای مثال آنچه در انتخابات ریاست جمهوری در سال ۲۰۱۶ در آمریکا رخ داد. محققان در تلاش هستند تا متوجه شوند که چگونه استفاده از داده‌های پنجاه میلیون کاربر Facebook توانست نتیجه انتخابات ریاست جمهوری آمریکا را تغییر دهد.

How artificial intelligence conquered democracy, Vyacheslav W Polonski, 20 July 2019, https: //www. independent. co. uk/news/long_reads/artificial_intelligence_democracy_elections_trump_brexit_clinton_a7883911. html

3. The rise of the deepfake and the threat to democracy, Simon Parkin, 20 July 2019, https: //www. theguardian. com/technology/ng_interactive/2019/jun/22/the_rise_of_the_deepfake_and_the_threat_to_democracy

۴. ر.ک: توحیدی و عبدخدائی، ۱۳۹۸.

آیا پیاده‌سازی اصول اخلاقی در دستگاه‌های مجهز به هوش‌مصنوعی، چیزی معادل اخلاق طراحان و سازندگان دستگاه‌های هوشمند است؟به عبارتی، آیا ملزم نمودن طراحان به پایبندی به یک سری اصول اخلاقی به‌عنوان مرام‌نامه‌ها، رفتارنامه‌ها و یا به تعبیری کدهای اخلاقی، سبب می‌شود که در طراحی ماشین اصول اخلاقی رعایت گردد؟ گرچه پایبندی طراحان به رفتارنامه‌های اخلاقی یا به تعبیری اخلاقِ پیش از طراحی لازم است، اما به نظر می‌رسد کافی نیست؛ چراکه آنچه از توانایی‌ها و عملکرد هوش‌مصنوعی در دسترس است، نشان می‌دهد که قضیه بسی پیچیده‌تر است. به عبارتی گاهی دستگاه‌های مجهز به هوش‌مصنوعی تصمیم‌سازی‌هایی می‌نمایند که حتی طراحان هم دلیل این پیشنهاد و تصمیم‌سازی را نمی‌دانند. به همین دلیل برخی بر این باورند که طراحان ماشین مانند والدینی هستند که یکی از عوامل تصمیم‌گیری‌های او به شمار می‌آیند، ولی نقش صددرصدی در تصمیم‌گیری فرزندشان ندارند. این بدین معناست که طراحان ماشین‌های هوشمند تنها محل بحث در مسئله اخلاق ماشین نیستند؛ بلکه خود ماشین هوشمند به‌عنوان یک عامل اخلاقی مستقل نیز باید مورد بحث قرار گیرد. به عبارتی شبکه گسترده‌ای از عاملین، طراحان و کاربر انسانی گرفته تا عوامل بازاریابی ماشین،[1] همگی در این تصمیم‌گیری و عملکرد ماشین دخیل هستند و البته خود ماشین هوشمند هم باید به‌عنوان یکی از حلقه‌های این شبکه مورد بررسی قرار گیرد.

حال سؤال اینجاست که آیا اساسا می‌توان ملاحظات اخلاقی را در ماشین‌های هوشمند پیاده‌سازی کرد یا خیر؟ اگر می‌توان پیاده‌سازی کرد چه مسیرهایی پیش‌رو قرار دارد؟

پیاده‌سازی اخلاق در ماشین‌های هوشمند از ابعاد مختلف مهندسی هوش‌مصنوعی، روانشناختی و فلسفی مورد توجه و ملاحظه قرار گرفته است. (Gips, 2011: 252) با رویکرد هوش‌مصنوعی، مهندسان کدها و الگوریتم‌های استدلال اخلاقی را براساس چارچوبی خاص در ماشین‌های هوشمند طراحی می‌کنند. بدین‌معنا که برای مثال با طراحی سیستم‌های پردازشگر هوشمند، هواپیماهای جنگنده خودران بتوانند میان افراد مختلف، افرادغیرنظامی رااز دشمن تشخیص داده و انسان‌های بیگناه را نکشند. بدین نحو که در نهایت امکان، نواقص و شرایط مورد نیاز برای پیاده‌سازی هر کدام از چارچوب‌ها و رویکردهای اخلاقی، مورد بررسی قرار می‌گیرد.[2]

از بعد روان‌شناختی، محققان بر نحوه شکل‌گیری و رشد استدلال اخلاقی در انسان تمرکز کرده

1. Marketers of the machine

۲. برای مطالعه بیشتر رجوع کنید به:

McLaren, 2011\ 297 - 315; Guarini, 2011\ 316 - 334; Mackworth, 2011\ 335 - 360; Bringsjord & others, 2008\ 361 - 374; Turilli, 2011\ 375 - 397; Pereira & Saptawijaya, 2011\ 398 - 421

و با مطالعه آن، درصددند تا شرایط مشابهی را در جهت یادگیریِ اخلاقیِ ماشین فراهم کنند.[۱] در این راستا نظریاتی ارائه شده است که مهم‌ترین آن الگوگیری از نحوه یادگیری در کودکان است.[۲] بدین‌معنا که همانطور که کودک از والدین و نزدیکان خویش عملکردهای گوناگون ازجمله عملکردهای اخلاقی را یاد می‌گیرد، ماشین‌های هوشمند نیز با استفاده از الگوهای یادگیری می‌توانند در معرض کاربر اخلاقی قرار گرفته و عملکرد اخلاقی را از او یادگرفته و پیاده می‌نمایند.[۳]

از بُعد فلسفی با دو منظر کلی بحث مورد مطالعه قرار می‌گیرد؛ نخست از منظر مسائل فرااخلاق و دیگری از بعد مسائل کاربست اخلاق هنجاری در ماشین‌های هوشمند. از منظر فرااخلاق، مهم‌ترین موضوع، بحث عاملیت اخلاقی ماشین است. همانطور که مشهودست، طراحان و کاربران مصنوعات تکنیکی همواره در حال ارزیابی این مسئله هستند که آیا ابزارهای ساخته شده، در تحقق اهدافی که برای آن طراحی شده‌اند به‌طور موفق عمل می‌کنند یا خیر؟ علاوه بر ارزیابی هنجارهایی که نقش تعیین کننده در عملکرد دستگاه ها در راستای رسیدن به اهداف ظاهری آن را دارند، هنجارهای اخلاقی نیز باید مورد توجه قرار گیرند. مقصود از هنجارهای اخلاقی، هنجارهایی است که نقش تعیین کننده در ارزش‌گذاری جامعه ایفا می‌کنند؛ بدین‌معنا که گاهی ارزشی را در جامعه القا کرده و یا اینکه سبب تغییر ارزش‌ها در جامعه می‌گردند. بنابراین، باید مورد توجه و ارزیابی قرار گیرند. به چنین ماشینی که ارزش‌های اخلاقی در آن لحاظ شده و دارای عملکرد اخلاقی است، عامل اخلاقی گفته می‌شود.

برای عاملیت اخلاقی در ماشین‌های هوشمند می‌توان سطوح و انواع مختلفی در نظر گرفت. بدین صورت که ماشین در سایه عاملیت اخلاقی، از لحاظ وضعیت درونی،[۴] مورد بررسی قرار گرفته و پرسش‌هایی از این قبیل مطرح گردد که: یک عامل برای آنکه اخلاقی محسوب شود، باید واجد چه شرایطی باشد؟ آیا ماشین را می‌توان یک عامل اخلاقی تام در نظر گرفت؟ به عبارتی آیا ماشین

---

۱. برای مطالعه بیشتر رجوع کنید به:

Dehghani & others, 2011\ 422 - 441; Danielson, 2011\ 442 - 450

۲. نگارنده در این راستا نیز مقاله ای در دست دارد. این رویکرد، تحت عنوان رویکرد پایین به بالا مورد بررسی قرار گرفته است.

۳. چگونگی نحوه یادگیری ماشین درمورد شناسائی اشیاء مختلف بررسی و آزمایش شده است، اما در مورد یادگیری اصول اخلاقی هنوز در مرحله نظریه است. بدیهی است که ماشین‌های کنونی شاید بتوانند یک رفتار خاص را به عنوان صرفا یک رفتار یاد بگیرند، اما تمایز میان رفتار اخلاقی با رفتار غیراخلاقی نظر به اخلاقی بودن آن تا زمان نگارش این مقاله هنوز ممکن نیست.

4. Inner status

دارای ملاک‌هایی از قبیل خودمختاری (Autonomy)، قصدمندی (Intentionality)، مسئولیت‌پذیری (Responsibility)، آگاهی (Consciousness)، شخص بودن (Selfness) و عاطفه‌مندی (Emotionality) است؟[۱]

یافته‌های مهندسی هوش‌مصنوعی تا زمان نگارش این مقاله نشان می‌دهد که نمی‌توان ماشین هوشمند را عامل تام اخلاقی درنظر گرفت. به این معنا که ماشین مانند انسان هویت اخلاقی داشته و بتواند در هر نقش یا موقعیت شغلی جایگزین مناسبی برای انسان باشد. اما به هرحال، اگر بر فرض در آینده ماشین‌های هوشمندی ظهور یابند که بتوانند نقش‌های حساسی مثل پلیس یا قاضی را در اختیار گیرند و به‌طور مستقل توانایی استدلال اخلاقی داشته باشند، یا به عبارتی عامل اخلاقی تام محسوب شوند، در آنصورت پرسش‌های گسترده‌ای مطرح خواهد شد؛ اینکه مسئولیت تصمیم‌سازی در ماشین متوجه چه کسی است؟ خود ماشین مسئول است یا طراح آن و یا کاربر ؟ در این سطح از خودمختاری و عاملیت اخلاقی، چالش‌های به وجود آمده در مورد هویت و تمامیت اخلاقی ماشین نیز مطرح می‌شود که مجال بررسی آن در این مقاله نیست. اما ازآنجایی‌که ارتباط وثیقی میان مباحث فرااخلاق و کاربست اخلاق در ماشین وجود دارد، لازم است پیش‌فرض نویسندگان در مورد بحث عاملیت اخلاقی ماشین مشخص شود. در این نوشتار ماشین هوشمند عامل تأثیرگذار اخلاقی محسوب می‌شود؛ یعنی عملکرد آن به‌گونه‌ای است که تأثیرات اخلاقی به بار می‌آورد. بنابراین لازم است تا ماشین به‌گونه‌ای طراحی گردد که ملزم به رعایت مجموعه اصول اخلاقی باشد تا در تصمیم‌سازی‌های خودمختار، بهترین عملکرد یا به تعبیری بهترین رفتار اخلاقی ممکن را انجام دهد.[۲] حال پرسش اینجاست که منظور از بهترین رفتار توسط ماشین چیست؟ چگونه می‌توان براساس بهترین رفتار و بهترین عملکرد، ماشین را به گونه‌ای طراحی کرد که به هدف اخلاقی نزدیک باشد؟

در راستای کاربست اخلاق در ماشین‌های هوشمند، دو و یا به تعبیری سه رویکرد مورد بحث و نظر است؛ رویکرد نخست رویکرد بالا به پایین بوده که در این مقاله مورد مطالعه، بررسی و نقد قرار خواهد گرفت. رویکرد دوم رویکرد پایین به بالا و رویکرد سوم رویکرد تلفیقی است که در این مقاله مجال سخن پیرامون آن نبوده و در آینده به آن پرداخته خواهد شد.

---

۱. برای مطالعه بیشتر رجوع کنید به:

Dennet, 1996 : 351 - 365; Moor, 2011: 13 - 20; Tonkens, 2009: 421 - 428; Himma, 2009: 19 - 29

۲. برای مطالعه بیشتر رجوع کنید به:

Grau, 2011: 451 - 463; Powers, 2011: 464 - 475; Anderson Anderson, 2011a : 476 - 494; Anderson Anderson & Armen, 2005: 149 - 155

ازاین‌رو، مقاله حاضر درصددست تا به پاسخ پرسش‌های زیر نزدیک گردد؛ چگونه می‌توان میان رویکرد اصل‌گرایی در اخلاق با رویکرد بالا به پایین در پیاده‌سازی در ماشین قرابت یافت؟ براساس رویکرد وظیفه‌گرایی در اخلاق و یا بر مبنای فایده‌گرایی اخلاقی؟ آیا رویکرد اصل‌گرایی در اخلاق دربردارنده تمامی ابعاد بحث می‌باشد و یا باناکارآمدی‌هایی مواجه خواهد بود؟

بدین‌منظور نخست رویکرد اصل‌گرایی در اخلاق کاربردی مطالعه گردیده و کاربست دو نظریه اخلاقی فایده‌گرا و وظیفه‌گرا براساس نظریه بالا به پایین تبیین و بررسی خواهد شد. در انتها نقدها و چالش‌های این دو نظریه اخلاقی در پیاده‌سازی در ماشین‌های هوشمند بیان می‌شود.[۱]

الف) مفهوم‌شناسی بحث

اخلاق کاربردی[۲] شاخه‌ای از مطالعات فلسفه اخلاق است که هدف آن بهره‌گیری از اصول فلسفی در راستای ارائه راهکار به منظور رفع چالش‌های اخلاقی در زمینه‌های کاربردی نظیر سیاست، بهداشت و سلامت، مهندسی، فناوری و نظایر آن است. (Beauchamp, 2003: 1) در این میان اخلاق هوش‌مصنوعی، شاخه‌ای از اخلاق کاربردی است که به بررسی امکان کاربست اخلاق و همچنین روش‌ها و موانع پیاده‌سازی اخلاق در ماشین‌های هوشمند می‌پردازد. هدف این شاخه مطالعاتی یافتن روشی در راستای عملکرد مسئولانه و اخلاقی سیستم‌های هوشمند است. (Anderson ,2011a: 476 - 494) بدین‌منظور، محققان با مطالعه نحوه تفکر اخلاقی انسان و با الگوگیری از نحوه عملکرد اخلاقی در انسان، درصددند تا الگویی برای استدلال اخلاقی استخراج کرده و در ماشین‌های هوشمند پیاده‌سازی کنند. (Gips, 2011: 244) بنابراین گام اول آن است که دریابیم انسان به‌عنوان یک عامل اخلاقی چگونه استدلال‌های اخلاقی خود را پیش می‌برد. آیا انسان برای آنکه عملی را به لحاظ اخلاقی درست یا نادرست بداند، تابع استدلال و دلیل است و یا اینکه صرفا براساس شهود اخلاقی عملی را درست و یا نادرست می‌داند (Ridge & McKeever, 2016: 1)؟ در این راستا دو نظریه مطرح است؛ اصل‌گرایی[۳] و نمونه‌گرایی.[۴] براساس اصل‌گرایی، تنها زمانی می‌توان عملی را اخلاقی یا غیر اخلاقی دانست که از پیش، اصل و معیاری برای اخلاقی بودن یا نبودن آن وجود

---

۱. رویکرد اصل‌گرایی در اخلاق کاربردی را می‌توان در دو رویکرد وظیفه‌گرایی و فایده گرایی یافت. رویکرد فضیلت گرایی اخلاقی از حیطه این بحث خارج است چرا که رویکرد فضیلت گرایی در اخلاق علاوه بر در نظر گرفتن اصول بر اخلاق عملی نیز تأکید دارد.

2. Applied ethics.
3. Principle based.
4. Case based.

داشته باشد. بنابراین استدلال‌های اخلاقی از اصول کلی اخلاقی شروع شده و به نمونه‌ها و مصادیق عملی تسری می‌یابند. (Harman, 2005: 45 ؛ McNaughton, 1988: 190 - 192) حال آنکه در نظریه نمونه‌گرایی، انسان‌ها براساس مصادیق و نمونه‌های عملی در شرایط خاص تصمیم اخلاقی اتخاذ می‌کنند. به عبارتی در تصمیم‌گیری‌های اخلاقی انسان، مصادیق اولویت دارند نه اصول از پیش‌تعیین شده و لذا بر مبنای مصادیق می‌توان یک اصول کلی تدوین نمود. (Dancy, 1993: 60)

از سوی دیگر، مطابق با دو نظریه اصل‌گرایی و نمونه‌گرایی، دو رویکرد متداول در پیاده‌سازی اخلاق در ماشین وجود دارد که عبارت‌اند از رویکرد بالا به پایین[۱] و پایین به بالا.[۲] رویکرد بالا به پایین در اخلاق ماشین عبارت است از تعیین مجموعه‌ای از قواعد و اصول کلی؛ اصولی که قابلیت تبدیل به یک الگوریتم را داشته باشند. بر این اساس، ابتدا اصول اخلاقی از پیش تعیین شده و سپس بر موقعیت‌های مختلف إعمال می‌شوند. در این رویکرد، آن‌چه یک عامل اخلاقی مصنوعی[۳] در جهت عمل اخلاقی باید انجام دهد، این است که اعمال خود را تحت آن قوانین محاسبه کند تا مشخص شود که آیا انجام چنین عملی براساس قوانین کلی اخلاقی مجاز است یا خیر. (Wallach & Allen, 2008: 79 & 84) حال آنکه، رویکرد پایین به بالا عبارت است طی کردن فرآیند تکامل در یادگیری به منظور رسیدن به عمل اخلاقی. بنابراین، رویکردهای پایین به بالا بر این باورست که برای داشتن یک ماشین اخلاقی نباید از اصول کلی شروع کرد؛ بلکه همچنان که ذهن انسان با توجه به شرایطی که از ابتدای کودکی با آنها در تعامل است، استدلال اخلاقی را کسب می‌کند، ماشین نیز با شبیه‌سازی این فرآیند می‌تواند به این مهارت دست یابد که در هر موقعیتْ عمل درست را تشخیص داده و به فراخور آن تصمیم‌گیری کند. (Allen, Smit & Wallach, 2005: 149 - 150)

همانطور که مشهود است، با تأمل در رویکرد اصل‌گرایی در اخلاق می‌توان مشابهت ساختاری میان این رویکرد و رویکرد بالا به پایین[۴] در طراحی و پیاده‌سازی الگوریتم‌های ماشین یافت؛ به عبارتی مطابق با رویکرد اصل‌گرایی در اخلاق، رویکرد بالا به پایین در اخلاق ماشین ارائه داد. پیاده‌سازی این رویکرد در ماشین، رؤیای قدیمی فلاسفه بوده تا بدین‌وسیله ماشین‌هایی تولید شوند که از انسان بهتر عمل می‌کنند. برای مثال لایب‌نیتس[۵]، فیلسوف آلمانی قرن هفده، با طراحی ماشین محاسبه‌گر به این امید بود که روزی ماشین‌های قوی‌تری بتوانند اصول کلی اخلاقی را در جهت

1. Top - down method.
2. Bottom - up method.
3. Artificial moral agent – AMA.
4. Top - down method.
5 . Leibniz

محاسبه بهترین عمل در هر شرایطی به کار ببندند (Ibid: 83).[۱]

سیستم‌های اخلاقی بالا به پایین لزوما دارای یک خاستگاه و یک سری اصول واحد نیستند؛ بلکه این خاستگاه می‌تواند دین (نقل)، فلسفه (عقل)، ادبیات (فرهنگ و عرف جامعه) و نظایر آن باشد؛ قاعده طلایی،[۲] ده فرمان،[۳] اخلاق نتیجه‌گرا یا فایده‌گرا،[۴] امر مطلق اخلاقی کانت،[۵] کدهای قانونی ـ حرفه‌ای و سه قانون رباتیکز آسیموف[۶] را می‌توان به‌عنوان مثال نام برد.

در میان مکاتب فلسفی که استدلال اخلاقی را تحت یک قاعده واحد کلی مطرح می‌کنند، دو مکتب فایده‌گرایی و وظیفه‌گرایی به نظر می‌رسد در ماشین قابل اجرا باشند. (Allen, Smit & Wallach, 2005: 150) در ادامه به اختصار این دو نظریه اخلاقی شرح داده شده و سپس کاربست آن در ماشین هوشمند بررسی خواهد شد.

ب) فایده‌گرایی[۷]

فایده‌گرایی یا سودگرایی یکی از نظریه‌های مهم اخلاقی در پاسخ به این پرسش است که «چگونه باید عمل کنیم؟» ازآنجاکه فایده‌گرایان بر نتیجه عمل تأکید دارند، این دیدگاه نوعی نتیجه‌گرایی اخلاقی[۸] نیز به حساب می‌آید.[۹] فایده یا منفعت[۱۰] با میزان خشنودی یا بهزیستی در ارتباط است. براساس این نظریه و در یک تقریر کلی، هدف نهایی اخلاق «به حداکثر رساندن مقدار ممکن سود یا منفعت در جهان» است. لذا اصل کلی فایده‌گرایی عبارت است از «آن گونه عمل کن که نتیجه یا پیامد عملت بتواند بیش‌ترین میزان ممکن منفعت را برای بیش‌ترین افراد ممکن فراهم آورد». (Frankena, 1988: 34)

---

۱. بر خلاف رویای لایب‌نیتس، امروزه از نظر دانشمندان اخلاق، چنین رویکردی در استدلال اخلاقی انسانی موفقیت آمیز عمل نمی‌کند، چراکه انسان‌ها از انجام تمامی محاسبات مورد نیاز در مورد سنجش رفتار خویش عاجزند.

2. The Golden Rule.

3. Ten Commandments.

4. Utilitarian Ethics.

5. Kant's Moral Imperative.

6. Asimov's Three Laws of Robotic.

7. Utilitarianism.

8. Moral consequentialism.

۹. رویکرد نتیجه‌گرایی اخلاقی بر نتایج حاصل از عملکرد انسان متمرکز است؛ بدین معنا که اگر نتیجه عملی خوب باشد، آن عمل درست بوده و اگر نتیجه حاصل از عمل بد باشد، آن عمل نادرست. حال پرسش آن است که منظور از خوب یا بد بودن نتیجه نسبت به کیست؟ از نظر ایگوئیسم اخلاقی، رفتاری درست یا مطلوب است که دارای بهترین نتیجه برای صاحب آن رفتار باشد. درحالیکه از منظر فایده‌گرایی اخلاقی، رفتاری درست است که دربردارنده حیطه وسیعی از خوبی‌ها برای تعداد زیادی از انسان‌ها باشد. مبحث ایگوئیسم اخلاقی مورد بحث این بخش از مقاله نیست.

10. Utility.

دو رویکرد فایده‌گرایی عبارتند از؛ فایده‌گرایی عمل‌محور و فایده‌گرایی قاعده‌محور. در رویکرد اول، عمل هر شخص در راستای تولید بیش‌ترین منفعت، مورد ارزیابی قرار می‌گیرد. اما در رویکرد دوم، قاعده‌ای که براساس آن رفتار صورت گرفته در راستای تولید بیش‌ترین منفعت، ارزیابی می‌شود. برای مثال اگر شخصی برای نجات جان انسان دیگر دروغ بگوید، براساس فایده‌گرایی عمل‌محور، آنچه اهمیت دارد و باید مورد ارزیابی قرار گیرد، عمل فرد در شرایط خاص است. اما براساس فایده‌گرایی قاعده‌محور، قاعده عمل، یعنی اصل دروغ‌گویی به‌طورکلی، مورد ارزیابی قرار می‌گیرد. (Ibid: 35 - 39)

دیدگاه فایده‌گرایی با جرمی بنتام[1] فیلسوف تجربه‌گرای انگلیسی شناخته می‌شود. بنتام به ایده «اخلاق محاسباتی»[2] (اخلاق براساس محاسبه) شهرت دارد. (Wallach & Allen, 2008: 86 و Anderson Anderson, 2011b: 476) براساس این دیدگاه، اخلاق براساس یک شالوده عینی بنا می‌گردد؛ بدین‌معنا که موقعیت‌های مختلف اخلاقی و سود و زیان آن به لحاظ کمّی قابل ارزیابی بوده و می‌توان به آنها اعدادی را نسبت و وزن داد.[3] وجود مقیاس‌های کمّی در محاسبه منفعت، یک قاعده ساده را پیش رو می‌گذارد: «عملی را انتخاب کن که بالاترین سود کلی را نتیجه دهد». در نتیجه شعار سنت فایده‌گرا این است: «بیش‌ترین خشنودی، برای بیش‌ترین تعداد» (Bentham, 1781: 31 - 34) در نظر گرفتن صرف ملاحظات کمّی در ارزیابی سود و زیان ناشی از اعمال، بعدها توسط جان استورات میل[4] فیلسوف انگلیسی مورد انتقاد، و ملاحظات کیفی نیز مورد توجه قرار گرفت. (Mill, 1863: 56)

در این رویکرد، هدف طراحی سیستم با کارکرد اخلاق محاسباتی است؛ کارکردی که به شیوه‌ای صحیح، منافع کنونی را نسبت به زیان‌های آینده بسنجد (و برعکس)، و یا بتواند به منافع و مضرات حقیقی نسبت به خطرات و سودهای بالقوه وزن دهد. بر این اساس، برای دست‌یابی به ملزومات محاسباتی، یک ماشین اخلاقی نتیجه‌گرا، باید واجد چهار توانایی باشد:[5]

---

1 . Jeremy Bentham.

2. Moral Arithmetic.

۳. بنتام یک فایده‌گرای عمل‌محور است و برای محاسبه کمّی سود و زیان، چهار ملاک اصلی معرفی می‌کند: شدت و ضعف، مدت، قطعیت و دوری و نزدیکی سود یا زیان. (Bentham, 1781: 31)

4 . John Stauart Mill

۵. به دست آوردن ملزومات محاسباتی در راستای طراحی یک ربات نتیجه‌گرا در سال ۱۹۹۵ توسط جیمز گیپس دانشمند علوم کامپیوتری در کالج بوستون آمریکا انجام شد. (Wallach & Allen, 2008: 86)

۱. روشی برای توصیف یک موقعیت در جهان.

۲. روشی برای تولید اعمال ممکن.

۳. ابزاری برای پیش‌بینی اوضاعی که ممکن است به‌عنوان نتیجه تصمیمات کنونی به وجود آیند.

۴. روشی برای ارزیابی یک موقعیت بر حسب خوب یا مطلوب بودن. (Gips, 2011: 245)

اگرچه لیست ارائه شده توانست چارچوب سودمندی برای تعیین شروط جزئی‌تر را ارائه دهد، اما از سوئی معیارهای استاندارد برای تبدیل شدن به یک الگوریتم را نداشت و از سوی دیگر ابهامات بسیاری را باقی گذاشت. برای مثال، توصیف یک موقعیت تا چه اندازه باید کامل و دقیق باشد؟ ماشین باید توانایی انجام چه نوع اعمالی را داشته باشد؟ چگونه ماشین هوشمند موقعیت‌هایی را که به لحاظ زمانی و مکانی با آن فاصله زیادی دارند، پیش‌بینی می‌کند؟ چگونه موقعیت‌های متفاوت براساس سود و زیان ارزیابی می‌شوند؟

به منظور بررسی دقیق‌تر سوالات بالا، به بررسی سایر شروط پرداخته خواهد شد تا چالش‌های پیش روی ماشین‌های فایده‌گرا شفاف‌تر گردد.

بررسی شرط اول

ماشین هوشمند اخلاقی نتیجه‌گرا باید برای توصیف یک موقعیت در جهان، روش خاصی داشته باشد. حال پرسش آن است که عناصر مرتبط با یک موقعیت خاص کدام‌اند؟ با در نظر گرفتن وسعت حوزه اخلاق، این عناصر می‌توانند بسیار متنوع باشند؛ از انسان‌ها و حیوانات گرفته تا تمامی اکوسیستم‌های موجود در جهان (البته با این قید که می‌توان به هرکدام وزنی متفاوتی داد). فارغ از اینکه چگونه باید این عناصر مرتبط را تعیین کرد، چالش بعدی تخمین داده مورد نیاز برای یک مسئله اخلاقی خاص و سوژه‌های مورد نظر آن است. برخی معتقدند که با وجود این حجم انبوه داده، یک «عامل جهانی»[۱] نیاز است که ترجیحات تمامی موجودات را در حافظه خود داشته، به آن وزن داده و این اطلاعات را در کنار هم ارزیابی نماید تا به تصمیم درست اخلاقی به‌معنای نفع حداکثری رهنمون شود.[۲] چنین عاملی باید در حافظه خود، یک نشانی وب[۳] مشخص داشته باشد تا بدین‌وسیله قابلیت دسترسی به اطلاعات مرتبط را پیدا کند. بنابراین وجود چنین ایده‌ای از اساس نادرست است. (Wallach & Allen, 2008: 88)

---

۱. «The World Agent» یک عامل همه‌دان مطلق که ترجیحات همه موجودات و میزان سود یا ضرری را که در نتیجه تصمیم در موقعیت اخلاقی دریافت می‌کنند، می‌داند.

۲. برنارد ویلیامز فیلسوف مطرح بریتانیایی، معتقد است که چنین موجودی باید در کنار علم مطلقی که دارد، یک ناظر بی‌طرف هم باشد. او این موجود را «عامل جهانی» می‌نامد. او در نهایت، امکان وجود چنین عاملی را منکر می‌شود. (Williams, 1985: 83)

3. URL.

بررسی شرط دوم

ماشین هوشمند اخلاقی برای تصمیم‌گیری در مورد انتخاب رفتار و اعمال خویش باید روش خاصی داشته باشد. این شرط نیز تحت تأثیر عناصر سازنده یک موقعیت قرار می‌گیرد. به‌عنوان مثال، آیا رفاه حال حیوانات بخشی از معادله تصمیم‌سازی به حساب می‌آید یا خیر؟ در آن صورت آیا خوردن گوشت حیوانات، به‌عنوان یک عمل اخلاقی توسط ماشین لحاظ می‌شود یا خیر؟ از این‌رو، هرچه تنوع اطلاعات مرتبط با حقایق اخلاقی بیش‌تر باشد، لازم است برنامه‌ریزی جزئی‌تری برای انتخاب‌های ممکن ماشین صورت پذیرد. (Ibid)

بررسی شرط سوم

ماشین هوشمند اخلاقی برای تخمین تأثیرات گسترده به وجود آمده از یک رفتار یا یک عمل، باید روش خاصی داشته باشد. در این راستا، طراح الگوریتم با پرسش‌های زیر مواجه است؛ نخست آنکه چه بازه‌ای از زمان مد نظر است؟ آیا مقصود تأثیرات گسترده در آینده نزدیک است یا در آینده دور را نیز شامل می‌شود؟ مقصود از آینده نزدیک چیست؟ چه تأثیراتی در آینده دور قابل چشم‌پوشی است؟ (Ibid)

بدیهی است که هر عملی تأثیرات اولیه‌ای را به‌وجود می‌آورد که حتما باید ارزش‌های اخلاقی این تأثیرات محاسبه شوند. اما هر عمل، تأثیرات ثانویه نامحدودی را نیز به دنبال خواهد داشت. اگر بازه زمانی آینده محدود نباشد، ماشین مجبور به محاسبه احتمالات بی‌پایان تأثیرات ثانویه هر رفتار خواهد شد و بدین‌ترتیب زمان زیادی از کارکرد پردازنده مرکزی[1] صرف برنامه‌هایی می‌شود که تمام تعاملات و برهم‌کنش‌های محتمل را بررسی کند. بعلاوه، تأثیرات ثانویه می‌توانند زنجیره‌ای از تأثیرات دور از دسترس محتمل را نیز بوجود آورند؛ چیزی شبیه «اثر پروانه‌ای».[2] مشکل دیگری که محاسبه تأثیرات آینده را مسئله‌ساز می‌کند، نقص اطلاعات است؛ چرا که همه داده‌ها در اختیار نیست و ممکن است وقایع پیش‌بینی نشده‌ای به وقوع بپیوندد. بدین‌منظور شاید بهترین روش، شیوه هواشناسان برای پیش‌بینی وضع هوا باشد؛ بدین ترتیب که از چندین رایانه محاسبه‌گر استفاده شده و در نهایت از نتایج حاصل، میانگین گرفته می‌شود. به‌طور مشابه، در محاسبه منفعت آینده نیز می‌توان رویکردهای چندگانه را در پیش‌بینی نتایج اعمال و رفتار به کار گرفت و به یک برآیند کلی دست یافت. (Ibid: 89)

علاوه برآن، اصول و قوانین اخلاقی استخراج شده نسبت به انسان‌های کنونی و با توجه به شرایط

1. CPU.

۲. «Butterfly effectt» طبق این مثال یک ضربه بال زدن پروانه‌ای در آسمان چین، می‌تواند چند هفته بعد بر شرایط آب و هوایی شمال آمریکا تاثیرگذار باشد. (Wallach & Allen, 2008: 88)

مکانی و زمانی و همچنین نگرش آنان شکل گرفته‌است. اگر بناست ماشینی به روش بالا به پایین و براساس اصول اخلاق نتیجه‌گرایی ساخته شود، در طراحی آن باید از اهمیت نتایج دوردست کاست. حتی لازم است، میزان اهمیت نتایج با میزان عدم قطعیت وقایع، نسبت مساوی داشته باشد. بدین‌معنا که هرقدر امکان تحقق کمتری وجود داشته باشد، از اهمیت نتایج حاصل از آن نیز کاسته شود. اما این مسئله ساده نیست؛ چراکه فرمول مشخصی وجود ندارد تا به‌وسیله آن فاصله زمانی و مکانی را به عدم قطعیت نسبت داد؛ برای مثال ممکن است برخی وقایعی که در سال آینده یا در مسافت ۵۰۰۰ کیلومتر دورتر از اینجا اتفاق می‌افتند، از حوادثی که در هفته آینده یا در فاصله ۱۰۰ متری اینجا اتفاق می‌افتند، قابل پیش‌بینی‌تر باشند.

بررسی شرط چهارم

ماشین هوشمند اخلاقی، یک موقعیت را بر حسب خوبی یا مطلوبیت آن ارزیابی می‌کند. همان‌طورکه گفته شد، فایده‌گرایان، در مورد اینکه آیا باید برای لذت‌های متنوع وزن‌های متفاوت قائل شد، اختلاف نظر دارند.[۱] برای یافتن فرمول به منظور وزن‌دهی به لذت افراد و سپس پیاده‌سازی آن در یک عامل اخلاقی مصنوعی، یک روش آن‌است که میزان رتبه‌دهی افراد به لذات مختلف را جمع‌آوری کرده و سپس با استفاده از داده‌ها که به شکل تصاعدی تنظیم شده‌است، تصمیمات و رفتار عامل اخلاقی مصنوعی شکل گیرد. بدیهی است که این روش کارساز نبوده و مشکلات جدی بر سر راه جمع‌آوری اطلاعات مربوط به رتبه‌دهی لذات وجود خواهد داشت.

پس از بررسی چهار شرط در طراحی ماشین فایده‌گرا، چالشی دیگر در مورد مسئله محاسبه حائز اهمیت است؛ اینکه محاسبه به‌طور سلسله‌وار ادامه پیدا کند و ماشین نتواند در موقعیت مناسب به موقع عمل نماید بدین‌معنا که به منظور دست‌یابی به تمام چهار شرط فوق، ماشین در پردازش بی‌پایان قرار گیرد یا آنکه محاسبه نتایج احتمالی یک تصمیم، خود یک عمل محسوب شده و از این‌رو، موقعیتی است که بر آن نتایج اخلاقی بار می‌شود و همین طور این سلسله ادامه می‌یابد؟ بنابراین، باید راه‌کار مناسبی برای پایان دادن محاسبه وجود داشته باشد. براساس فایده‌گرایی، اگر عمل محاسبه در میزان منفعت حاصل شده، تأثیر منفی داشته باشد، باید به‌عنوان یک عمل غیراخلاقی شناخته و متوقف گردد. برای مثال اگر شخصی فرصت کمک به یک نیازمند را به‌خاطر طولانی شدن فرآیند تصمیم‌گیری از دست بدهد، آن‌گاه فرآیند تصمیم‌گیری کارکرد صحیحی نخواهد داشت چراکه در نهایت به تصمیم به‌موقع و صحیح منجر نشده است. اما سوال اینجاست که

---

۱. برای مثال میل در این رابطه با بنتام مخالف است. از نظر او کیفیت لذات باید در محاسبه ارزش اعمال به عنوان یک معیار دخیل باشد. (Mill, 1863: 14)

بدون اینکه تمام ابعاد و جوانب عملی محاسبه گردد، چگونه می‌توان فهمید که آن عمل ارزش اخلاقی دارد یا خیر؟ برای حل این مشکل می‌توان پیشنهاد داد که توقف فرآیند طولانی محاسبه ماشین، توسط همان عامل اخلاقی مصنوعی صورت نگیرد، بلکه از ماشین دیگری کمک گرفته شود بدین معناکه در هنگام پردازش و محاسبه موقعیت‌های گوناگون، به یک‌باره ماشین از ارزیابی متوقف گردد. اما پرسش هنوز باقی است که آیا این روش، یعنی تصمیم‌گیری بدون ارزیابی ابعاد گوناگون پیامدهای یک رفتار، موجب عملکرد اخلاقی در ماشین می‌شود؟

برخی از نظریه‌پردازان[۱] بر این باورند که فایده‌گرایی برای تصمیم‌گیری‌های اخلاقی نمی‌تواند یک نظریه سودمند و عملی باشد چراکه حتی انسان نیز با این قبیل چالش‌ها در تصمیم‌گیری مواجه است.[۲] بنابراین همانطور که انسان‌ها برای تصمیم‌گیری از «عقلانیت محدود»[۳] استفاده می‌کنند، عامل اخلاقی مصنوعی نیز به همان ترتیب عمل نماید. عقلانیت محدود، شامل مجموعه محدودی از ملاحظاتی است که در تصمیم‌گیری عقلانی فرد لحاظ می‌شود. حال سوال این است که آیا یک سیستم محاسباتی محدود، چیزی که دقیقا مانند انسان عمل کند، همان چیزی است که ما از ساخت یک عامل اخلاقی انتظار داریم؟ آیا این یک فاعل اخلاقی کارآمد خواهد بود؟ آیا ممکن نیست که ماشین‌ها به دلیل اطلاعات محدود خود، برآورد نادرستی از نتایج دور و نزدیک رفتار خود داشته باشند؛ آنچه موجب به وجود آمدن درد و رنجی است که بشر در طول تاریخ، از جانب فاعل‌های انسانی، شاهد آن بوده است؟

برای گریز از این مشکل، روش «جستجوی مکاشفه‌ای»[۴] پیشنهاد شد؛ منطقی که به جای جستجو در میان تمام حالات ممکن، فرآیند جستجوی پیچیده در سیستم‌های هوشمند را کوتاه‌تر می‌نمود.[۵] روش مکاشفه‌ای در اخلاق نیز قابل توسعه است. در یک نوع روش مکاشفه‌ای اخلاقی در

---

۱. هِرب سایمون پدر علم هوش مصنوعی و برنده جایزه نوبل اقتصاد در سال ۱۹۸۴، نظریه عقلانیت محدود را مطرح کرد. (Simon, 1997)

۲. در واکنش به این دیدگاه می‌توان گفت، انسان‌ها در تصمیم‌گیری‌های خود ملاحظات فایده‌گرایانه دارند، هرچند که آن را به نحو ایده آل انجام نمی‌دهند؛ مثلا گاهی عملی انجام می‌دهند با این نیت که توسط آن، میزان آسایش و رفاه جمعی را به حداکثر برسانند اما در عین حال علم مطلق به همه چیز ندارند و انتظار هم ندارند که در آن موقعیت همه چیز را بدانند.

3. Bounded Rationality.

4. Heuristic Search.

۵. این روش که توسط سایمون و همکارش آلِن نیوول ارائه گردید، نقش اساسی در موفقیت سیستم‌های هوشمندی نظیر Deep Blue II ایفا کرد. بدین ترتیب که، Deep Blue، ربات شطرنج باز شرکت IBM، برای انجام بازی شطرنج دیگر لازم نبود تا در میان تمام حرکات ممکن در یک زمان نامشخص به جستجو بپردازد. در عوض، با استفاده از روش‌های تقریبی تخمین، روی دستیابی به اهداف میان مدت تمرکز می‌کرد: قاعده‌ای که چیدمان خاصی از مهره‌ها را روی صفحه شطرنج، ارزشمندتر از بقیه چیدمان‌های ممکن تخمین می‌زند. (Wallach & Allen, 2008: 90)

سیستم‌های فایده‌گرا، تنها باید نتایج و پیامدهای مستقیم هر عمل رتبه‌بندی شود و پیامدهای ثانویه اخلاقی مورد توجه نیست. مقصود از پیامدهای ثانویه، درجه ارتباط یک عمل با پیامدهای اخلاقی آن است؛ برای مثال، اگرچه سرنگونی یک دولت خارجی، نتایج بلندمدتی برای سیاست‌مداران به بار می‌آورد و قطعا نیازمند بررسی و ارزیابی است، اما تأثیرات این عمل بر برنامه‌کاری روزنامه‌نگاران، مسئله‌ای نیست که نیاز به توجه داشته باشد. (Ibid: 90)

در نوع دیگری ازاین‌روش تنها قواعدی بررسی می‌شوند که با منافع جمع محدودی مرتبط هستند، بدین‌ترتیب دایره بررسی محدودتر می‌گردد. هرچند بررسی محدود در راستای بهینه سازی، در بسیاری از مواقع می‌تواند موجب بهینه‌سازی عمومی نیز گردد؛ برای مثال فرض کنید که شخصی در حال ارزیابی و تحلیل نتایج عمل خویش است و برای این کار تنها نتایج عمل را در جهت افزایش منفعت جامعه کوچک خود لحاظ می‌کند. در این شرایط احتمال بیش‌تری وجود دارد که یک جامعه سالم، به هنگام نیاز جوامع دیگر به آنها کمک کند و این احتمال برای یک جامعه ناسالم کم‌تر است؛ به‌علاوه یک فاعل اخلاقی، ممکن است بدون اینکه مجبور باشد تمامی این روابط را بررسی کند، به‌طور موثر در زنجیره تأثیرات نقش داشته باشد. (Ibid: 90 - 91)

به‌طور خلاصه می‌توان گفت که لحاظ جنبه‌های محاسباتی در نظریه فایده‌گرایی، در نگاه اولیه این دیدگاه را مناسب‌ترین گزینه برای پیاده‌سازی در ماشین هوشمند نشان می‌دهد؛ چراکه ماشین هوشمند سریع‌تر و دقیق‌تر از انسان، نتیجه یا منفعت حاصل از یک عمل را محاسبه می‌کند. اما چنان‌که شرح آن رفت، کارکرد محاسباتی سودها و زیان‌های اعمال در ماشین با چالش‌های جدی مواجه است؛ نحوه ارزیابی عناصر مرتبط با یک عمل، در نظر گرفتن دامنه انتخاب‌های ممکن ماشین نسبت به زمان، مکان و افراد، تخمین میزان تأثیرات ناشی از عمل در بازه زمانی مشخص ازجمله چالش‌هایی هستند که یک ماشین فایده‌گرا با آن مواجه است. (Powers, 2011: 464) این مسائل از آنجا ناشی می‌شوند که معیار سنجش اخلاقی بودن یا نبودن یک عمل، عینی و دارای مصادیق زیادی است؛ به نحوی که می‌تواند در هر عمل و در مورد هر قاعده محاسبات متفاوتی را طلب کند. بعلاوه نظریه فایده‌گرایی به‌عنوان یک دیدگاه موجه اخلاقی با انتقادهای جدی مواجه است. ازجمله این انتقادها، ناعادلانه بودن نحوه سنجش درستی اعمال است؛ به گونه‌ای که هدف (بالا بردن سود)، وسیله را توجیه می‌کند. (Grau, 2011: 453 - 454) بنابراین در شیوه پیاده‌سازی رویکرد اخلاقی بالا به پایین در ماشین، باید دیدگاهی مورد مطالعه قرار گیرد که فارغ از محاسبه سود و زیان نتیجه اعمال، به‌طور مشخص‌تری وظایف اخلاقی را در نظر گیرد.

دیدگاه وظیفه‌گرایی در اخلاق ازجمله نظریات اخلاقی است که به جای تمرکز بر روی نتیجه اعمال، به عمل اخلاقی به خودی خود و فارغ از لحاظ نتیجه آن تاکید دارد. در ادامه امکان و نحوه پیاده‌سازی این دیدگاه در ماشین بررسی می‌شود.

ج) وظیفه‌گرایی

وظیفه‌گرایی یکی دیگر از نظریات غالب در میان رویکردهای بالا به پایین است. بر خلاف دیدگاه‌های نتیجه‌گرا یا فایده‌گرا که ملاک اخلاقی بودن عمل را نتیجه عمل و یا فایده آن می‌دانند، نظریه‌های وظیفه‌گرا در تعیین اصول اخلاقی، نتیجه عمل را لحاظ نمی‌کنند، بلکه آنچه اولیت دارد، انجام یک سری وظایف و الزامات اخلاقی است. نظریه وظیفه‌گرا مجموعه‌ای از اصول است که راهنمای عمل قرار داده شده‌اند، فارغ از اینکه این اصول نتایج مثبتی برای شخص یا جامعه هم‌نوع او دربردارد یا خیر. از این‌رو، تأکید آن به خوبی و بدی ذاتی اعمال است، نه نتایج مفید یا مضر آن. (Spielthenner, 2005: 221 - 222) قواعد اخلاقی وظیفه‌گرا، می‌توانند بسیار صریح بوده و به‌طور خاص در مورد یک رفتار در یک موقعیت باشند؛ مثل اینکه «نباید دزدی کنی». و یا اینکه این قواعد و اصول انتزاعی بوده، بیانگر قاعده کلی باشند و از آنها قواعد جزئی‌تر به دست آید؛ مثل اینکه «با دیگران طوری رفتار کن که دوست داری همان‌گونه با تو رفتار شود». ده فرمان کتاب مقدس و سه قانون اخلاقی آسیموف[1] ازجمله نمونه‌های جزئی‌تر وظیفه‌گرایی هستند و برای نمونه‌های انتزاعی‌تر، می‌توان نسخه‌های متفاوتی از قانون طلایی که در بسیاری از ادیان و فرهنگ‌ها به اشکال مختلف وجود دارد و یا امر مطلق اخلاقی کانت را نام برد.

در این بخش، قواعد رباتیکز آسیموف به‌عنوان یک نمونه جزئی وظیفه‌گرا[2] و امر مطلق اخلاقی کانت به‌عنوان یک نمونه انتزاعی وظیفه گرا، به ترتیب بررسی شده و چالش‌های پیش‌روی آنها در طراحی ماشین مطرح می‌گردد.

هرجا بحثی از رویکردهای بالا به پایین اخلاقی به میان می‌آید، سه قانون رباتیکز آسیموف مورد بحث قرار می‌گیرند.[3] سه اصل اخلاقی آسیموف عبارت‌اند از:

۱. ربات نباید به‌طور مستقیم یا تعمدی و یا به‌طور غیرمستقیم و در تعامل با انسان موجب آسیب‌زدن به انسان شود.

---

1. Asimov.

۲. برخی قواعد آسیموف را نمونه‌ای از اخلاق فایده‌گرا دانسته و به بحث در مورد امکان و چالش‌های پیاده‌سازی آن در ماشین هوشمند پرداخته‌اند. (برای مثال نگاه کنید به: Grau, 2011\ 451 – 461)

۳. کلارک(Clarke) و آندرسن (Anderson)به طور مفصل به بحث در مورد کاربست قواعد آسیموف در ماشین هوشمند پرداخته‌اند. (نگاه کنید به Clarke, 2011: 254 - 284 و Anderson 2011: 285 - 296)

۲. ربات باید از دستورات انسان اطاعت کند، مگر زمانی که این دستورات با قانون اول در تعارض باشند.

۳. ربات باید از خودش محافظت کند، مگر زمانی که این محافظت با قانون اول یا دوم در تعارض باشد. (Asimov, 1968: 26)

بعد از آنکه آسیموف این سه قانون اخلاقی را تدوین کرد، یک قانون چهارم یا صفرمی را اضافه کرد. (چون جایگزین سه قانون قبلی شد، این‌گونه نامیده شد)

قانون صفرم: یک ربات نباید به انسان صدمه بزند یا به واسطه عدم اقدام خود، موجب شود که به انسانی صدمه برسد. (Asimov, 1985: 373)

ایده‌ای که آسیموف در مورد عامل اخلاقی مصنوعی مطرح کرد، این بودکه قواعد اخلاقی ماشین‌های هوشمند بیش‌تر از آنکه تابع قواعد انسانی باشند، باید مطابق استانداردهای متفاوت دیگری باشند. به عبارت دیگر، آنچه در طراحی قوانین آسیموف مدنظر بود، آنست که ربات‌ها دارای هویتی همچون بَرده و در خدمت اهداف بشر باشند و اصول اخلاقی تنها در سایه حکمرانی انسان مورد دغدغه است.[۱] (Gips, 2011: 244) فارغ از این چالش، اگرچه اصول اخلاقی آسیموف در نگاه اول، سر راست به نظر می‌رسند، ولی اعتبار آنان در تبدیل شدن به یک الگوریتم، چنان‌که توضیح داده خواهد شد، مورد تردید است.

چالش اول

مشکل اولی که ماشین با آن مواجه است، خطا در درک مفاهیم موجود در قواعد و تعمیم آنهاست؛ مثلاً آیا با توجه به قواعد اخلاقی آسیموف، رباتی که تنها قادر به درک معنای مستقیم واژه‌ها است، باید مانع قطع اعضای بدن بیمار توسط جراح شود؟ اطمینان از این که ربات درک کند که جراح، با قطع اعضای بدن بیمار قصد آسیب رساندن به او را ندارد، کار ساده‌ای نیست. (Clarke, 2011: 260) پیاده‌سازی یک سیستم ساده اخلاق وظیفه‌گرا، مستلزم ایجاد درکی از بافت فرهنگی ـ زبانی و همچنین در نظر گرفتن استثناها برای این قواعد است؛ قواعد جزئی که نشان دهند در چه استثنائاتی یک عمل خاص، صدمه به انسان محسوب نمی‌شود. چنین ماشین اخلاقی هوشمندی، نیاز به دانش و اطلاعات وسیعی دارد تا بتواند قواعد کلی را به درستی در بافت‌ها و موقعیت‌های متفاوت شناسائی و به‌کار گیرد، و همچنین دانش آن براساس یافته‌ها و موقعیت‌های جدید به روز شود. (Wallach & Allen, 2008\ 92)

---

۱. برخی قواعد آسیموف را از این چالش مبرا می‌دانند. برای مثال به عقیده گراو «Grau» ازآنجاکه ربات‌ها دارای آگاهی، خودمختاری و شخصیت نیستند، پیاده‌سازی اصولی اخلاقی که صرفا بر منفعت انسان تاکید دارد، ضد اخلاق نیست. (Grau, 2011: 458)

چالش دوم

مشکل اصلی قواعد وظیفه‌محور، امکان تعارض قواعد اخلاقی با یکدیگر است. آسیموف این مشکل را با اولویت‌بندی قواعد اخلاقی حل کرد؛ بدین صورت‌که اگر عملی از جانب انسان به ماشین امر شده باشد و در عین حال موجب آسیب به انسانی دیگر شود، قانون دوم به نفع قانون اول کنار می‌رود؛ زیرا صدمه نرساندن به انسان از اطاعت او برتر است. اما حل این مسئله به این سادگی نیست؛ حتی یک قاعده واحد به تنهایی می‌تواند موجب تعارض در تصمیم‌گیری شود. مثل اینکه دو انسان، دو دستور متناقض به یک ماشین بدهند. (Clarke, 2011: 264) این مسئله باعث توقف کار ماشین شده و تصمیم‌گیری را غیرممکن می‌کند. در این مواقع، هر عامل اخلاقی مصنوعی قاعده‌محور، نیازمند قاعده درجه بالاتری است تا موقعیت‌هایی را که در آن قواعد با هم تعارض دارند، مدیریت کند.

با تمام این اوصاف، در واکنش به سناریوی وظیفه‌محور آسیموف، بسیاری از متخصصان علوم کامپیوتری به این نتیجه رسیده‌اند که برای ساخت یک عامل اخلاقی مصنوعی وظیفه‌گرا، باید قواعد انتزاعی را در نظر گرفت؛ قواعدی که قدرت انعطاف بیش‌تری در موقعیت‌های خاص اخلاقی داشته باشند و تعارضات میان قواعد جزئی با ارجاع به آنها رفع شود. (Wallach & Allen, 2008: 94 - 95)

امر مطلق کانت، نمونه قاعده انتزاعی است که به‌طور خاص برای تضمین انسجام منطقی میان قواعد طراحی شده است؛ بنابراین برای سیستم‌هایی که از چارچوب منطقی پیروی می‌کنند، کاربرد خواهد داشت. (Ibid: 95) نسخه اصلی این قاعده عبارت است از: «همواره تحت قاعده‌ای عمل کنید که در عین حال بتوانید اراده کنید که آن قاعده به یک قانون جهان شمول تبدیل شود». از همین قاعده انتزاعی، قاعده‌های جزئی‌تر نیز قابل استخراج است. مثلاً اگر شما ملزم به اجرای این قاعده باشید، نمی‌توانید زیر قول خود بزنید. زیرا اگر بخواهید قاعده عمل خود، یعنی زیر قول زدن را، به شکل قاعده کلی در جهان درآورید، در آنصورت همه زیر قول خود می‌زنند. آن‌گاه قول دادن از اساس کارکرد خود را از دست خواهد داد. (Kant, 1785: 15)

به‌طور کلی، براساس نظریه کانت، یک سیستم وظیفه‌محور باید دارای شرایط زیر باشد:

۱. هدف از انجام عمل خود را تشخیص دهد.

۲. تأثیرات اعمال دیگر فاعل‌های اخلاقی را که در تلاش برای رسیدن به همان هدف با همان شیوه هستند، ارزیابی کند.

۳. در ابتدایی‌ترین حالت بداند که باید چگونه عمل کند؛ زیرا تنها چیزی که (۱) و (۲) می‌گویند این است که یک ماشین چگونه عمل نکند.

۴. دانش وسیعی از روان‌شناسی انسانی داشته باشد تا بتواند عمل خود را در جهت رضایت انسان‌هایی که در تعامل با آنهاست، ارزیابی کند. (Wallach & Allen, 2008: 96)

حال پرسش اینجاست که ماشین چگونه می‌تواند امر مطلق را شناسائی کرده و براساس قاعده کلی عمل نماید؟ اولین حدس معقول این است که: رباتی که قرار است در میان گزینه‌های مختلف برای عمل، یکی را انتخاب کند، ابتدا باید بررسی کند که اگر عامل‌های دیگر در شرایط او قرار داشته و تصمیم مشابه او را می‌گرفتند، آیا باز هم هدف او قابل تحقق بود یا خیر؟ تعیین چنین اصلی که براساس آن بتوان موقعیت‌هایی مختلف را شناسائی کرد و براساس قاعده کلی ارزیابی نمود، بسیار دشوار و شاید غیرممکن باشد؛ چراکه به دانش انتزاعی در مورد اهداف، اعمال، شرایط و محیط نیاز خواهد بود. بعلاوه، چنین ماشینی باید تا حد زیادی به روان‌شناسی انسان و ربات و همچنین به تأثیرات رفتارها بر یکدیگر آگاه باشد.

علاوه بر آن، رسالت قواعد انتزاعی این بود که تعارض میان اصول جزئی را رفع کنند، حال آنکه چالش هنوز باقی است؛ چراکه مسئله سازگاری قواعد کلی استخراج شده با یکدیگر و همچنین تشخیص اینکه چه قواعدی با قواعد سطح بالا سازگاری دارند، خود کاری پیچیده و دشواری است. همچنین مشکل ابهام مفاهیم که در سیستم‌های اخلاقی هوشمند وظیفه‌گرا بیان شد، در این قواعد نیز وجود دارد. همان‌طور که گفته شد، عامل اخلاقی باید درک درستی از قواعد داشته باشد تا در موقعیت‌های اخلاقی استدلال کرده و عمل نماید. اگر قواعد کلی، دستوری را واضح و شفاف ارائه می‌داد، آن‌گاه تفسیر و کاربرد آن آسان بود؛ حال آنکه همیشه چنین نیست؛ زیرا برای آنکه قواعد کلی، کاملا شفاف بیان شوند، سیستم هوشمند باید تعاریف واضحی از عباراتی که در این قواعد استفاده می‌شود را داشته باشد و ایجاد درک دقیق از این تعاریف کار ساده‌ای نیست. برای مثال چگونه ماشین می‌تواند درک شفافی از واژه «جهان شمول» در امر مطلق کانت داشته باشد؟ و یا اینکه چگونه ماشین به‌طور مشخص مفهوم صدمه یا آسیب به انسان را درک خواهد کرد؟ در این راستا برخی راه حل‌های گوناگونی را برای پیاده سازی اخلاق کانتی در ماشین پیشنهاد داده و بررسی کرده‌اند؛ راه حل‌هایی نظیر «سازگاری محض»، «استدلال عملی بر مبنای عقاید عمومی» و «ارتباط و وابستگی» (خزاعی و زمان فشمی، ۱۳۹۲) در هر کدام از این مراحل تلاش بر این بوده تا از نگرش سنتی کانتی کوتاه آمده و نگرشی منعطف به کانت ارائه گردد تا آنجا که ماشین به «پایگاهی از اطلاعات شخصی» مبدل می‌شود. برخی نیز بر این باورند که این مفاهیم در مقام تعریف مشکل‌ساز هستند ولی همین مفاهیم مبهم، در عمل می‌توانند کاربرد شفافی داشته باشند؛ به‌طور مثال، مفهوم «کچلی» مبهم است و نمی‌توان گفت یک شخص در ازای کمبود چند تار مو کچل محسوب می‌شود، ولی در هر صورت بدون داشتن یک تعریف مشخص همه ما

می‌دانیم که شخصی مثل دالای‌لاما[1] کچل است. به‌طور مشابه ممکن است مفهوم «آسیب» خیلی شفاف نباشد، اما به وضوح می‌توان فهمید که برخی از افعال آسیب‌زا هستند. بنابراین، به‌نظر می‌رسد با تمرکز بر مثال‌های کاربردی واضح و شفاف، بسیاری از ابهامات در رویکرد اخلاقی از بالا به پایین رفع شده و در نهایت سیستم‌های هوشمند این قابلیت را کسب خواهند کرد تا بر این اساس استدلال کرده و عمل نمایند. علاوه برآن، عامل مصنوعی که تحت قاعده وظیفه‌گرایی طراحی می‌شود، باید قواعد (یا چگونگی تعیین قواعد) را دانسته و برای به کار بستن قواعد در موقعیت‌های خاص اخلاقی روشمند باشد. همچنین عامل اخلاقی باید بتواند به‌طور منسجم قواعد جزئی و خاص را اعتبار کند. طراحان ماشین اخلاقی وظیفه‌گرا، باید اطمینان حاصل کنند که ماشین در موقعیت‌های مختلف، قواعد مرتبط اخلاقی را اتخاذ کرده، آنها را به درستی به کار بسته، ابهام در قواعد را رفع کرده و در هنگام بروز تعارض قواعد با یکدیگر، آنها را مدیریت می‌کند. (Wallach & Allen, 2008: 96 - 97)

د) بررسی و ارزیابی

گذشته از نکاتی که در خلال بحث به آنها اشاره شد، نکته قابل تأمل و کلی آن‌که اساس و بنیان قوانین اخلاقی کانت بر مبنای عامل اخلاقی انسانی بنا گردیده است؛ بنابراین بحث از توان‌مندی‌های درونی انسان نظیر عقلانیت، اراده، اختیار و استقلال رأی نقش به‌سزایی در جایگاه اخلاقی کانت ایفا می‌نماید. اگرچه هدف این نوشتار بحث از جایگاه عاملیت اخلاقی نبوده و تنها هدف بررسی امکان کاربست اخلاق از منظر وظیفه‌گرایی اخلاقی در ماشین است، اما نادیده گرفتن این بنیان فکری سزاوار نیست. بنابراین پرسش اصلی باقی مانده آن است که آیا ماشین قادر به استخراج قوانین الزام‌آور جهان شمول به تنهایی می‌باشد یا آن‌که نیازمند هدف‌گذاری و دخالت انسانی در این راستاست؟ در هر دو حالت این چالش اخلاقی مطرح خواهد بود که در نهایت چه کسی قانون‌گذار بوده و چه کسی تصمیم‌گیرنده نهایی در یک رفتار اخلاقی است؟ آیا تصمیم‌گیری‌های اخلاقی ماشین کاملا تحت کنترل بوده و در شرایط متفاوت رفتارهای متفاوتی اعمال خواهد گردید؟ این مسئله زمانی پیچیده‌تر خواهد شد که حتی در صورت هدف‌گذاری اصول توسط انسان برای ماشین، ماشین قوانین جدیدی را استخراج نموده و وظایف خویش را براساس قوانین جدید خود که نه قابل توضیح‌پذیری است، نه قابل پیش‌بینی و نه قابل قطعیت، اعمال نماید. آیا ماشین قادر است قوانین جهان شمولی را که ذاتاً مطلوب بوده و ارزشمند است، استخراج کند؟

مفهوم ارزش به خودی خود در جوامع مختلف مفاهیم گسترده‌ای دارد. فهم چگونگی

1. DalaiLama.

قضاوت‌های اخلاقی براساس اصولی که بسیاری از آنها انتزاعی هستند، کار ساده‌ای نیست. اگرچه چارچوب‌های اخلاقی متعددی پیشنهاد شده است، اما درعین حال بر سر یک چارچوب خاص، توافق عام وجود ندارد. همین مسئله سبب می‌شود که طراحان ماشین‌های هوشمند با ابهام، کلیت و تنوع اصول اخلاقی مواجه شوند؛ تا آنجا که استدلال اخلاقی را «حوزه تحلیلی ضعیفی»[1] می‌دانند که به دشواری می‌توان به آن پرداخت. مفهومی، انتزاعی و نسبی بودن مصادیق اخلاقی سبب گردیده که همواره مباحث اخلاقی به صورت «زمینه ـ باز»[2] مطرح گردد؛ بدین‌معنا که در زمینه‌های مختلف تصمیم‌گیری، می‌توانند معانی متعددی داشته باشند، تا حد زیادی در معرض تفسیرهای گوناگون قرار گیرند و به اندازه کافی دقیق نباشند تا بتوان در مورد هر موقعیت اخلاقی آنها را به کار بست. علاوه بر آن، تعارض قواعد با یکدیگر از ویژگی‌های دیگر این حوزه است، به گونه‌ای که نمی‌توان هیچ راه‌حل جامعی برای گریز از آن در پیش گرفت. (McLaren, 2011: 297 - 311)

همچنین فایده‌گرایی بنتام حتی توسط هم‌کیشان او نیز به دلایل مختلف اخلاقی مورد انتقاد قرار گرفت؛ اینکه ملاک و معیار نفع عمومی چیست؟ چگونه ماشین هوشمند قادر به تشخیص بیشترین نفع عمومی است؟ آیا ملاک حق و حقیقت انتفاع است و یا میزان رضایت عمومی؟ اما به‌طور کلی پرسش اصلی مطرح آن است که آیا اخلاق قادر به پیاده‌سازی محاسباتی است؟ آیا استدلال‌ها و اصول اخلاقی را می‌توان به صورت چکیده و خلاصه درآورد و با دانش تجربی و با اعداد و ارقام پیاده‌سازی کرد؟ همانطورکه گفته شد انسان‌ها برای آنکه تصمیماتی اتخاذ نمایند، یک‌سری وقایع را درنظر می‌آورند. در دنیای حقیقی ارزش‌ها با وقایع و اتفاقات توأم گردیده و فهم انسان را شکل می‌دهد. به عبارتی اخلاق یک امر شناختی است. از نظر دانشمندان، کاربست اخلاق در ماشین زمانی میسر است که ماشین بتواند اولا قوانین بی‌طرفانه و دارای جایگزین را اجرا نماید؛ به عبارتی در شرایط مختلف تصمیمات مختلف اتخاذ نماید و در ثانی بتواند تشخیص دهد که از چه طریقی باید این قوانین را اعمال نماید تا بیشترین نفع عمومی را دربرداشته باشد. در یک جمله بدون یافتن راه حل شناختی برای اعمال اخلاقی در ماشین تصور کاربست اخلاق نگاهی خوش‌بینانه است. (Moor, 1995)

اما از سوی دیگر، با وجود تمام مواردی که گفته شد، راه حل چیست؟ آیا می‌توان ماشینی را که رفتارهای هوشمندانه از خود نشان می‌دهد، در تمام ابعاد زندگی انسان دخیل شده است، انبوهی از داده‌های انسان را در اختیار دارد، در بسیاری از موارد به جای او تصمیم می‌گیرد و راهکار پیشنهاد می‌دهد، بدون سازوکار اخلاقی رها نمود؟ به بهانه نسبی بودن مصادیق اخلاق، آیا می‌توان از کلیات

1. Weak analytic domain.
2. Open – textured.

و قوانین اخلاقی چشم‌پوشی کرد؟ چگونه مسیری قادر است در عین در نظر گرفتن مصادیق جزئی کاربست اخلاق و انتخاب عملکرد اخلاقی به حسب شرایط مختلف، در عین حال پایبند به اصول و کلیات اخلاقی باشد؟ بنابراین درنظر گرفتن رویکردی که در عین آنکه وضع اصول کلی را لازم می‌داند، از مصادیق اخلاقی چشم پوشی نمی‌کند ضروری به نظر می‌آید.

نتیجه

دیدگاه فایده‌گرایی در اخلاق با آن‌که در نگاه اول به سبب ماهیت محاسباتی خود برای پیاده‌سازی در ماشین مناسب به نظر می‌رسید، اما با چالش‌های محاسباتی متعددی همراه است. نحوه ارزیابی عناصر مرتبط با یک عمل، در نظر گرفتن دامنه انتخاب‌های ممکن ماشین نسبت به زمان، مکان و افراد، تخمین میزان تأثیرات ناشی از عمل در بازه زمانی مشخص، ازجمله چالش‌های ذکر شده در پیاده‌سازی فایده‌گرایی در ماشین هوشمند هستند. برای گریز از این چالش‌ها و هم چنین به جهت آنکه اصول اخلاقی به‌طور مشخص‌تری ارائه شوند، دیدگاه وظیفه‌گرایی برای پیاده‌سازی در ماشین مطرح شد. اما دیدگاه وظیفه‌گرایی با مسئله تعارض در قواعد اخلاقی جزئی مواجه است. برای پرهیز از مشکل تعارض، پیاده‌سازی قواعد کلی‌تر اخلاقی پشنهاد می‌شود. چالش اصلی این دسته از قواعد ابهام در آنها و همچنین فقدان شیوه‌ای مناسب برای استخراج قواعد جزئی‌تر از قواعد کلی است. با توجه به مطالب گفته شده، به نظر می‌رسد که دیدگاه بالا به پایین به تنهایی برای یافتن یک رویکرد جامع در جهت پیاده‌سازی اخلاق در ماشین کافی نیست و در نظر داشتن دیدگاه‌هایی که تربیت ماشین را براساس موقعیت‌های اخلاقی ضروری می‌دانند، لازم خواهد بود. ازاین‌رو، رویکردهای دیگری ازجمله رویکرد پایین به بالا پیشنهاد می‌گردد که براساس آن یادگیری رفتارهای اخلاقی در کاربست اخلاق در ماشین لحاظ می‌گردد و یا رویکرد تلفیقی بیان می‌شود که به عبارتی ترکیبی از هردو رویکرد می‌باشد. بنابراین، بررسی امکان کاربست و مطالعه چالش‌های پیش‌روی سایر رویکردها در مورد پیاده‌سازی اخلاق در ماشین‌های هوشمند، در تحقیقات دیگر به‌طور مفصل بررسی خواهد شد.

منابع و مآخذ

۱. توحیدی، ن. عبدخدائی، ز.، ۱۳۹۸، «زمستان‌دوم در اخلاق هوش مصنوعی: آیا می‌توان برای پیشگیری تدبیری اندیشید؟»، همایش ***هوش مصنوعی و محاسبات نرم‌افزار در علوم انسانی***، دانشگاه علامه طباطبائی.

۲. خزاعی، ز. زمان‌فشمی، ن، ۱۳۹۲، «روش‌ها و موانع پیاده سازی اخلاق کانتی در ماشین‌های هوشمند»، ***پژوهش‌های اخلاقی***، شماره ۱۳، ۵ ـ ۲۲.

3. Allen, C., Smit, I., & Wallach, W, 2005, "Artificial morality: Top - down, bottom - up, and hybrid approaches". Ethics and information technology, 7 (3) , 149 – 155.
4. Anderson, M., Anderson, S., & Armen, C, 2005, "Towards machine ethics: Implementing two action - based ethical theories". Paper presented at the Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics.
5. Anderson, M., & Anderson, S. L, 2011) a. Machine ethics. New York: Cambridge University Press.
6. Anderson, S. L, 2011, The Unacceptability of Asimov's three laws of robotics as a basis for machine ethics. Machine Ethics, Cambridge University Press, Cambridge (UK) , 285 – 296.
7. Anderson, S. L., & Anderson, M, 2011) b. A prima facie duty approach to machine ethics: Machine learning of features of ethical dilemmas, prima facie duties, and decision principles through a dialogue with ethicists. Machine Ethics, Cambridge University Press, Cambridge (UK) , 476 – 494.
8. Asimov, I, 1985, Robots and Empire. In: London: Harper Collins.
9. Asimov, I, 1968, I, robot. London: Grafton Books.
10. Beauchamp, T. L, 2003, The nature of applied ethics. A companion to applied ethics, 1 – 16, Bringsjord, 2008.
11. Bentham, J, 1781, An introduction to the principles of morals and legislation. Clarendon Press.
12. Bringsjord, S, 2008, "Ethical robots: the future can heed us". In Ai & Society, 539 – 550.
13. Clarke, R, 2011, "Asimov's laws of robotics: implications for information technology". In Machine Ethics, Cambridge University Press, Cambridge (UK) , 254 – 284.
14. Dancy, J, 1993, Moral reasons. Wiley – Blackwell.
15. Danielson, P, 2011, "Prototyping N - reasons: a computer mediated ethics machine". In Machine Ethics, Cambridge University Press, Cambridge (UK) , 442 – 450.
16. Dehghani, M., Tomai, E., Forbus, K. D., & Klenk, M, 2011, "An Integrated Reasoning Approach to Moral Decision - Making". In Machine Ethics, Cambridge University Press, Cambridge (UK) , 422 – 441.
17. Dennett, D, 2006, "Computers as prostheses for the imagination". In Invited talk presented at the International Computers and Philosophy Conference, Laval, France, May, Vol. 3.
18. Dennett, D, 1996, "When Hal Kills, Who's to Blame?" In D. Stork (Ed.) , Hal's Legacy (pp. 351 - 365, Cambridge, MA: MIT Press.
19. Frankena, W. K, 1988, Ethics. Pearson; 2nd edition.
20. Gips, J, 2011, "Toward the ethical robot". In Machine Ethics, Cambridge University Press, Cambridge (UK) 244 – 253.
21. Grau, C, 2011, "There is no 'I' in 'Robot': robots and utilitarianism". In Machine Ethics, Cambridge University Press, Cambridge (UK) , 451 – 463.
22. Guarini, M, 2011, "Computational neural modeling and the philosophy of ethics reflections on the particularism - generalism debate". In Machine Ethics, Cambridge University Press, Cambridge (UK) , 316 – 334.

23. Harman, G, 2005, “Moral particularism and transduction”. Philosophical Issues, 15, 44 – 55.
24. 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.
25. Kant, I, 1785, Groundwork of the Metaphysics of Morals. Mary Gregor translation. In: Cambridge: Cambridge University Press.
26. Mackworth, A. K, 2011, “Architectures and ethics for robots: constraint satisfaction as a unitary design framework”. Machine Ethics, Cambridge University Press, Cambridge (UK) , 335 – 360.
27. McLaren, B. M, 2011, “Computational models of ethical reasoning: Challenges, initial steps, and future directions”. Machine Ethics, Cambridge University Press, Cambridge (UK) , 297 – 315.
28. McNaughton, D, 1988, Moral vision: An introduction to ethics. Wiley – Blackwell.
29. Mill, J. S, 1863, Utilitarianism. London: Parker, Son and Bourn.
30. Moor, J. H, 1995, “Is Ethics Computational”, Metaphilosophy, Vol. 26, No. 1/2 (January/April 1995) , pp. 1 - 21.
31. Moor, J. H, 2006, “The nature, importance, and difficulty of machine ethics”. In Machine Ethics, Cambridge University Press, Cambridge (UK) , 13 – 20.
32. Newell, A., & Simon, H. A, 1975, “Computer science as empirical inquiry: Symbols and search”. Communications of the ACM, 113 – 126.
33. Pereira, L. M., & Saptawijaya, A, 2011, Modelling morality with prospective logic. Machine Ethics, Cambridge University Press, Cambridge (UK) , 398 – 421.
34. Powers, T. M, 2011, “Prospects for a Kantian machine”. In Machine Ethics, Cambridge University Press, Cambridge (UK) , 464 – 475.
35. Ridge, M., & McKeever, S, 2016, “Moral particularism and moral generalism. ” On Stanford. Edu.
36. Simon, H. A, 1997, “Models of bounded rationality: Empirically grounded economic reason”, Vol. 3, MIT press.
37. Spielthenner, G, 2005, “Consequentialism or deontology?” Philosophia, 33 (1) , 217 – 235.
38. Sullins, J. P, 2006, “When is a robot a moral agent?” in Machine Ethics, Cambridge University Press, Cambridge (UK) , 151 – 160.
39. Tonkens, R. S, 2009, “Ethical implementation: A challenge for machine ethics”. In 2nd Symposium on Computing and Philosophy.
40. Turilli, M, 2011, “Ethical protocols design”. In Machine Ethics, Cambridge University Press, Cambridge (UK) , 375 – 397.
41. Wallach, W., & Allen, C, 2008, Moral machines: Teaching robots right from wrong. Oxford University Press.
42. Williams, B, 1985, Ethics and the Limits of Philosophy. Cambridge: Harvard University Press.