ریاضی و جامعه
شاپا (چاپی): ۲۳۴۵-۶۴۹۳، شاپا(الکترونیکی): ۶۵۰۷ – ۲۳۴۵
جلد ۷، شماره ۲ (۱۴۰۱)، صص. ۳۷-۶۷
http://www.ui.ac.ir
http://math-sci.ui.ac.ir

# منطق اخلاقی و اخلاق منطقی تعامل انسان و ماشین در موج سوم هوش مصنوعی

فریا نصیری‌مفخّم*، غزال محسنی و آرمین جعفرپیشه

**چکیده.** علم و زبان منطق، زیربنای بسیاری از رویه‌های استنتاجی و استدلالی در هوش مصنوعی برای مدل‌سازی تفکر و تصمیم‌گیری مغز انسان بوده است. از جمله تصمیم‌گیری‌هایی که حتی برای خود انسان‌ها نیز بغرنج است، تشخیص درست و نادرست در موقعیت‌های اخلاقی و اجتماعی–حقوقی است. تصمیم‌ها در این موقعیت‌ها متفاوت و برگرفته از مکاتب و نظریه‌های اخلاقی مختلف هستند. پیاده‌سازی این‌گونه تفکر در ماشین‌های هوشمند، این موضوع را چالش برانگیزتر نموده است. نیاز به پیاده‌سازی اخلاق در ماشین‌های هوشمند به‌دلیل تصمیم‌ها و اقداماتی است که به‌صورت خودکار و بدون دخالت انسان انجام می‌دهند. هراس از تسلط سامانه‌های خودمختار بر آینده بشر، باعث ظهور شاخه‌ای به نام اخلاق در هوش مصنوعی شده است که به اخلاقی بودن تعامل انسان–ماشین می‌پردازد. برای اعتماد به ربات‌ها، آن‌ها باید بتوانند چرایی و چگونگی تصمیماتی که اتخاذ کرده یا می‌کنند را به انسان‌ها توضیح دهند. این مطالعه، مفاهیم اخلاق ماشینی و ماشین اخلاقی را به‌همراه ربات نرم‌افزاری دارای اخلاق مبتنی بر منطق، به‌عنوان قدم‌هایی که پژوهش‌ها برای اطمینان از سامانه‌های هوشمند، مستقل و هم‌زیست با انسان‌ها توجه کرده‌اند، مرور می‌کند. این بررسی برای مثال‌های کلاسیک شامل واگن فراری، قایق، و دروغگویی و براساس اصول اخلاقی شامل فایده‌گرایی، پارتو، و اثر مضاعف، توصیف می‌شود. همچنین، مقایسه عملکرد چنین سامانه هوشمند در قالب یک ماشین خودران با تصمیماتی که انسان‌هایی از کشورها و فرهنگ‌های مختلف در این موقعیت‌های لبه اخلاقی اتخاذ می‌کرده‌اند نشان داده می‌شود.

## ۱. مقدمه

اخلاق هوش مصنوعی که در چند سال اخیر به عنوان شاخه‌ای داغ در هوش مصنوعی ظهور کرده است، به‌دنبال امتداد بخشی به بهره‌مندی بشر از خدمات فناوری‌های هوشمند و خودآیین علی‌رغم ترس‌ها و نگرانی‌هایی است که از هوش مصنوعی برای آینده بشر مطرح شده است [۲۰]. این شاخه از هوش مصنوعی به بررسی، طراحی و پیاده‌سازی عامل‌های هوشمند و خودمختار به‌عنوان ماشین‌ها و سامانه‌هایی می‌پردازد که در تعامل با انسان‌ها (و سایر ماشین‌ها) توانایی تصمیم‌گیری اخلاقی و قابل قبول از نظر انسان‌ها را داشته و اخلاق‌مدارانه رفتار نمایند. با وجود نوظهور بودن این شاخه از هوش مصنوعی، پژوهش‌های پرشماری به پیاده‌سازی چنین تفکری پرداخته‌اند و در نیمی از آن‌ها، استدلال‌های مبتنی بر منطق مورد استفاده بوده است [۲۱].

این مقاله به این مقدمه مختصر بسنده کرده، و به منظور ترویج و معرفی این شاخه، در بخش‌های ۲ و ۳، گذری بر اخلاق ماشینی و ماشین اخلاقی دارد. بخش ۲، موج‌هایی که هوش مصنوعی از آغاز تا به امروز و برای آینده سپری کرده است و نیاز

عبارات و کلمات کلیدي: تعامل انسان و ماشین، تعامل اخلاقی انسان و ربات، هوش مصنوعی توضیح‌پذیر، اخلاق کاربردی، اخلاق ماشینی، ماشین اخلاقی، منطق، معماهای اخلاقی، اتومبیل خودران.
دبیرتخصصی رابط: صغری نوبختیان
نوع مقاله: ترویجی
تاریخ دریافت: ۱۴۰۱/۰۶/۰۶ تاریخ پذیرش: ۱۴۰۱/۰۸/۱۷
* نویسنده مسئول
. http://dx.doi.org/10.22108/MSCI.2022.134930.1528

به شفافیت تصمیم‌ها و هوش مصنوعی توضیح‌پذیر[۱] را بیان می‌کند. بخش ۳، به‌قواعد اخلاقی مور[۲]، پیاده‌سازی نظریه‌های اخلاقی که مورد توجه پژوهش‌گران بوده‌اند، و قوانین رباتیک اخلاقی آسیموف[۳] می‌پردازد. اغلب مطالب این دو بخش از کتاب «هوش سفید: از اخلاق ماشینی تا ماشین اخلاقی» [۲۶] اخذ شده است. یکی از کارهای بنیادین در راستای پیاده‌سازی اخلاق ماشینی و ماشین اخلاقی، کتابخانه نرم‌افزاری مبتنی بر منطق به نام هیرا[۴] برای پیاده‌سازی ربات دارای اخلاق به نام ایمانوئل[۵] است که توسط آزمایشگاه اصول هوش مصنوعی دانشگاه فرایبورگ آلمان[۶] و دانشگاه صنعتی دانمارک[۷] ارائه شده است [۱۴]. بخش ۴، به‌طور کامل به‌توصیف آن پژوهش اختصاص دارد. در بخش ۵، تصمیم‌گیری میلیون‌ها انسان از کشورها و فرهنگ‌های مختلف در موقعیت‌های تصمیم‌گیری اخلاقی، و تصمیمی که ماشین مبتنی بر اخلاق در آن موقعیت‌ها می‌گرفته است، به نقل از چند پژوهش اخیر مقایسه شده است. جمع‌بندی و چند پیشنهاد در بخش ۶ می‌آید.

## ۲. مقدمه‌ای بر اخلاق ماشینی

خدمات هوش مصنوعی، عرصه‌های مختلفی از زندگی بشر را دربرگرفته است، مانند تشخیص بیماری‌ها، تعیین فرمول شیمیایی داورها، اعتبار و بیمه، نظارت و ممیزی، اقتصاد و تجارت بین‌المللی، امور دفاعی و نظامی، وسایل نقلیه خودران[۸] و اثبات جرم یا بی‌گناهی و صدور احکام کیفری. تصمیم‌هایی که سامانه‌های هوشمند در برهم‌کنش با انسان‌ها اتخاذ می‌کنند، مستقیم و یا غیرمستقیم بر زندگی انسان‌ها تأثیر دارد. این سامانه‌ها می‌توانند منجر به برخی پیامدهای مغرضانه یا خصمانه، به‌ویژه سوگیری‌های ناشی از داده‌های آموزشی آن‌ها (برخاسته از تعصب‌های نژادی، جنسیتی، صنفی و همانند آن) بشود. قابل فهم بودن عملکرد، درک‌پذیری الگوریتم یادگیری آن‌ها، تفسیرپذیری[۹]، و توضیح‌پذیری[۱۰] تصمیم‌های آن‌ها برای اطمینان از بی‌طرفی و فارغ از سوگیری بودن پردازش‌ها و توصیه‌های آن‌ها، از موانع اصلی هوش مصنوعی در اجرای عملی آن به‌عنوان یک «هوش عام مصنوعی[۱۱]» بوده است [۴، ۱، ۱۵، ۹، ۲۵، ۲۶].

هوش مصنوعی با سپری‌کردن بهار آغازین در دهه ۵۰ میلادی و زمستان آن در دهه ۷۰ میلادی و تابستان‌ها و زمستان‌های متعدد، درخششی دوباره و چندین موج را تجربه کرده است. دانشمندان هوش مصنوعی در موج اول (دانش دست‌ساز[۱۲]) و موج دوم (یادگیری آماری[۱۳]) لازم بود که برنامه‌نویس[۱۴] و متخصص داده‌ها[۱۵] باشند. در موج اول، هوش ماشین برگرفته از بینش برنامه‌نویسِ خود برای حل مشکلات خاص بوده است. به‌عنوان مثال، سامانه‌های خبره[۱۶] مبتنی بر دانشِ دامنه هستند ولی هر کدام برای مسئله خاصی عمل می‌کنند و برای مسئله دیگر، پاسخگو نیستند. مثلاً سامانه‌ای که برای تشخیص انواعی از بیماری‌های چشم است [۲۷]، قادر به تشخیص بیماری‌های خون نیست [۸]. با این‌وجود، سامانه‌های خبره به‌خوبی مجهز به توانایی پاسخگویی به سؤالاتی همچون «چه»، «چرا»، و «چگونه» بوده‌اند که دلیل پاسخی که به کاربر داده‌اند را به وی نشان دهند. امکان بررسی رویه استنتاج، پاسخ را برای کاربر مقبول‌تر می‌کند [۸، ۱۸]. موج دوم، یعنی یادگیری آماری با انقلابی در ۲۰۱۰ به نام یادگیری عمیق[۱۷] بر آن بود که خودِ سامانه‌های هوش مصنوعی مدل‌هایی بسازند و توضیح دهند که پدیده‌های دنیا چگونه کار می‌کنند، ولی برای ایجاد هوش عام مصنوعی که توانایی یادگیری، تفکر، و حل مسئله هم‌سطح انسان داشته باشد ناکارامد بوده است. یک مدل یادگیری ماشینی، همبستگی داده‌هایی که از آن‌ها می‌آموزد را کشف می‌کند ولی الزاماً در شفاف نمودن رابطه علت و معلولی و تفسیر

[1]XAI (eXplainable Artificial Intelligence) [2]Moor [3]Assimov [4]HERA (Hybrid Ethical Reasoning Agents) [5]IMMANUEL (Interactive Moral Machine bAsed oN mUltiple Ethical principLes) [6]Foundations of Artificial Intelligence Lab, University of Freiburg, Germany [7]Danish Technical University, Denmark [8]self-driving cars [9]interpretability [10]explainability [11]artificial general intelligence [12]handcrafted knowledge [13]statistical learning [14]programmer [15]data scientist [16]expert systems [17]deep learning

جعبه سیاه خود توانمند نیست. برای مثال، باورهای طراح یا کاربر در برهم‌کنش با ماشین، بر آموزش و یادگیری آن اثر می‌گذارد و بنابراین تصمیم‌های آن سوگیرانه و نادرست خواهد بود [۲۶].

در موج سوم هوش مصنوعی که در چند سال اخیر در راه بوده است، سامانه‌های مستقلی لازم است که بتوانند منطق خود را تفسیر کنند و توضیح دهند، چرا که پایبندی این سامانه‌ها به ارزش‌های اخلاقی، اجتماعی و حقوقی از ملزومات هوشمند بودن، استقلال و هم‌زیست بودن آن‌ها در جوامع بشری شده است. آن‌ها باید بتوانند برای تشخیصی که داده‌اند استدلال خود را به‌صورت قانع‌کننده توضیح دهند. این توضیح‌پذیری برنامه‌های کاربردی هوشمند لازم است که برای توجیه، کنترل، دریافتن، و بهبود رفتار آن‌ها صورت گیرد؛ یعنی انسان در برهم‌کنش با یک ماشین هوشمند، به جای درماندن در پرسش‌هایی همچون «چرا آن کار را انجام دادی؟»، «چرا کار دیگری انجام ندادی؟»، «چه موقع به نتیجه می‌رسی؟»، «چه مواقعی به نتیجه نمی‌رسی؟»، «چه موقع می‌توانم به تو اعتماد کنم؟»، و «چگونه می‌توانم خطاهای پیش‌آمده را تصحیح نمایم؟»، باید به این یافته‌ها نائل شود: «می‌فهمم که چرا آن کار را انجام دادی»، «می‌فهمم که چرا کار دیگری انجام ندادی»، «می‌دانم که چه موقع به نتیجه می‌رسی»، «می‌دانم که چه مواقعی به نتیجه نمی‌رسی»، «می‌دانم که چه موقع می‌توانم به تو اعتماد کنم»، و «می‌دانم که چگونه می‌توانم خطاهای پیش‌آمده را تصحیح نمایم». یعنی، یک هوش مصنوعی توضیح‌پذیر لازم است که اعتماد، اطمینان، شفافیت و آگاهی‌دهندگی داشته باشد. شفافیت، بنا به «بخش سیاست‌گذاری شرکت آلن تورینگ» اصل بنیادین اخلاقی در هوش مصنوعی است [۲۶]. شفاف بودن، به‌قابلیت اعتماد به‌عملکرد سامانه‌های هوشمند می‌افزاید. در این موج از هوش مصنوعی، متخصصان هوش مصنوعی که سامانه‌های با توان یادگیری محتوایی[۱۸] ایجاد می‌کنند باید با اخلاق[۱۹] باشند به‌طوری‌که این سامانه‌ها در برهم‌کنش با انسان به صورت جعبه‌ای شیشه‌ای باشند؛ یعنی از یادگیری ماشینی متخاصم به یادگیری ماشینی عادل، پاسخگو و شفاف[۲۰] حرکت کنند. امروزه دانش هوش مصنوعی تنها به دنبال ساختن ماشین‌هایی با عملکرد انسان‌گونه نیست، بلکه پیاده‌سازی ساختار درونی ذهن انسان و چگونگی اتخاذ تصمیمات آن به‌گونه‌ای که پاسخگو و اخلاق‌مدار باشد دارای اهمیت شده است [۹، ۱۵، ۴، ۱، ۲۶].

شفافیت هوش مصنوعی در انجام اقدام مورد نظر از جنبه چگونگی روند پردازش سامانه هوشمند برای انجام اقدام مورد نظر، و توجیه چگونگی حصول آن اقدام از آن پردازش مطرح است. این شفافیت، سه وظیفه در قالب توجیه پردازش (منصفانه، امن و عاری از تبعیض)، توضیح محتوا و خروجی معنادار به لحاظ اجتماعی و به زبان روزمره برهم‌کنش‌های انسانی (چرا و چگونه در اتخاذ تصمیم یا بروز رفتار آن گونه عمل کرده است)، و توجیه خروجی (موجه، عاری از تبعیض، منصفانه، و ایمن) بر عهده طراحان پروژه‌های هوش مصنوعی قرار داده است. یعنی، سامانه هوشمند باید بتواند توضیحی مناسب از روند تصمیم‌گیری ارائه دهد به طوری که تصمیمات اتخاذ شده برای انسان قابل فهم باشد. پاسخ‌گویی به معنی تعهد به آشکارسازی، توضیح و توجیه رفتار، روش انجام مسئولیت‌ها و امور مالی با هدف تضمین کاربرد صحیح منابع، حفظ حقوق و ارزش‌ها، و تبعیت از قانون است. سامانه‌های هوشمند که به اعمال قدرت بر عموم مردم و حکمرانی مربوط می‌شوند، باید حتی از سطح پاسخگویی بالاتر نیز برخوردار باشند [۲۶].

سلاح‌های کشنده هوشمند خودمختار، دستکاری افکار اجتماعی، تسلط ماشین بر داده‌ها و نقض حریم خصوصی، تلفیق چهره افراد در سایر تصاویر، تبعیض‌آمیز بودن تصمیمات ماشین خودران، و بسیاری مسائل دیگر، ترس از تسلط سامانه‌های خودآیین (خودمختار، خودکار، خودگردان)[۲۱] هوشمند بر آینده بشریت و نگرانی‌هایی را از پیامدهای اخلاقی این سامانه‌ها در میان دانشمندان ایجاد کرده است؛ از جمله، این‌که [۲۶]:

- آیا می‌توان اخلاق را در ماشین‌های هوشمند پیاده‌سازی کرد و چگونه؟
- چگونه می‌توان مطمئن شد که ماشین اخلاقی رفتار می‌کند؟
- مسائل فرا اخلاق و عاملیت اخلاقی ماشین هوشمند چگونه و متوجه چه فردی است؟

[18] contextual learning [19] moral [20] FAT (F: Fair, A: Accountable, T: Transparent) ML [21] autonomous systems


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

خودگردانی[۲۲] ماشین‌ها (عامل‌ها، بات‌ها و ربات‌ها) می‌تواند درجاتی متفاوت داشته باشد [۲۶]:

(۱) تمام اقدامات ماشین توسط انسان کنترل شود،
(۲) اقدامات کلی توسط انسان مشخص شود و برای هر اقدام، خود ماشین تصمیم می‌گیرد،
(۳) وظیفه و هدف توسط انسان تعیین می‌شود و خود ماشین اقدامات لازم برای انجام وظیفه را مدیریت می‌کند،
(۴) خود ماشین و بدون دخالت انسان، اهداف را تعریف کرده و آن وظایف را انجام می‌دهد.

از آنجاکه نافرمانی ربات‌ها و آسیب رساندن آن‌ها به انسان‌ها از نگرانی‌های بالقوه‌ی هوش مصنوعی است، اخلاق ماشین به‌تازگی به‌عنوان یک زیرشاخه از هوش مصنوعی ظهور پیدا کرده است و در مورد تضمین رفتارهای اخلاقی عامل‌های هوشمند به‌بحث و تبادل نظر می‌پردازد. نباید از فرصت استفاده از فناوری‌های نوین سامانه‌های هوشمند به‌دلیل امکان بروز خطرات ناشی از آن محروم ماند؛ بلکه به‌منظور افزایش اعتماد عمومی، معیارهای مطلوب جامعه مبتنی بر ارزش‌های اجتماعی و فرهنگی باید در طراحی و کاربری سامانه‌های هوشمند رعایت شود. افزایش رو به رشد سامانه‌های پیچیده و هوشمند، اخلاق ماشین را برای بیشینه‌سازی فواید و کمینه‌سازی مخاطرات برخاسته از فناوری‌های جدید برای جامعه ضروری نموده است. در این عصر، هدف از خلق هوش مصنوعی تنها بهینه‌سازی زندگی انسان و برنده شدن به هر بهایی نیست؛ انسان علاوه بر خود باید به منافع اجتماعی دیگران نیز بیندیشد. افول اخلاق و انسانیت، زمستان بعدی هوش مصنوعی را خواهد آورد اگر به هوش مصنوعی حساس به ارزش[۲۳]، هوش مصنوعی مسئول[۲۴]، و هوش مصنوعی قابل اطمینان[۲۵] توجه نشود. اخلاق مصنوعی یا اخلاق ماشین (یا همان توانایی ماشین در تصمیم‌گیری اخلاقی) در صدد تضمین رفتار اخلاق مدارانه در ماشین است؛ یعنی تضمین کند که رفتار ماشین از لحاظ اخلاقی برای انسان قابل قبول است [۲۶، ۱].

## ۳. مقدمه‌ای بر ماشین اخلاقی

در برهم‌کنش انسان و ماشین[۲۶] دو نکته اساسی مطرح می‌شود [۲۶]:

- بر اساس چه معیاری تصمیم ماشین هوشمند درست یا نادرست است؟
- مسئولیت نادرستی تصمیم‌های ماشین و جبران خسارات ناشی از آن‌ها بر عهده چه کسی است؟

در تصمیم‌ها و اقدامات وسایل خودران، افراد (سرنشینان، عابران، قضات) باید بدانند که آیا تصمیم‌های اتخاذ شده توسط الگوریتم آن وسائل اخلاقی هستند؟ و اگر هستند تا چه اندازه؟ در هلند و لیتوانی، ماشین‌های خودران اجازه دارند بدون راننده انسانی در مسیرهای عمومی تردد کنند. رانندگان اتومبیل‌های هوشمند در آلمان با رعایت دستورالعمل‌های ترافیک جاده‌ای کشور (مصوب کمیسیون اخلاقی حمل‌ونقل و زیرساخت دیجیتالی در سال ۲۰۱۷) به‌شرح زیر می‌توانند راندن و کنترل خودرو را به سامانه راننده خودکار بسپارند [۱۳، ۲۶]:

(۱) حفاظت از اشخاص بر همه ملاحظات دیگر اولویت دارد.
(۲) در هنگام خطر، حفاظت از جان انسان، بر حیوانات و اموال ارجحیت دارد.
(۳) در شرایط بروز تصادفات اجتناب‌ناپذیر، هر نوع تمایز (سن، جنسیت، وضعیت جسمی یا روانی، مقام، و غیره) در مورد اشخاص در معرض اصابت، ممنوع است.
(۴) مسئولیت خسارت توسط سامانه راننده خودمختار به عهده شخص حقیقی یا حقوقی عهده‌دار سایر مسئولیت‌های این محصول است.
(۵) ماشین باید بتواند در شرایط اضطراری بدون دخالت انسان به حالت امن برود.

[22]autonomy [23]Value Sensitive AI [24]Responsible AI [25]Trustworthy AI [26]Human-Machine Interaction

از آنجا که حتی با وجود مقاصد نیک طراحان، هوش مصنوعی ممکن است باعث مخاطره و ضرر بشود، اتحادیه اروپا در سال ۲۰۱۸ با تأکید بر دارا بودن سه ویژگی قانونی، اخلاقی، و درست بودن، سند «هوش مصنوعی قابل‌اطمینان» را راهکاری برای پایبندی هوش مصنوعی به اصول و ارزش‌های اخلاقی می‌داند. این ارزش‌ها عبارتند از [۱۳، ۲۶]:

(۱) احترام به آزادی انسان،
(۲) بازداشتن از هر نوع آسیب،
(۳) رعایت انصاف،
(۴) توضیح‌پذیری.

توضیح‌پذیری، یکی از الزامات قانونی اتحادیه اروپا برای اعمال شفافیت در تصمیم‌گیری سامانه‌های خودمختار شده است.

همچنین اتخاذ رویکرد ترتیبی و سلسله مراتبی در الگوریتم تصمیم‌گیری سامانه هوشمند دارای اهمیت است. برای نمونه، در تصمیم‌های یک ماشین خودران نخست باید کاهش تلفات، سپس کاهش آسیب‌ها و بعد کاهش تأثیرات مخرب زیست‌محیطی، و در آخر کاهش زمان رانندگی مورد توجه قرار گیرد. همه تصمیمات اتخاذ شده در مورد ترجیح یک قاعده بر قاعده‌ای دیگر باید مستدل و مستند باشد. اگر تعارضات میان قواعد به‌صورتی باشد که نتوان یکی را بر دیگری ترجیح داد، سامانه هوشمند در هر مرحله‌ای (از ایجاد تا استفاده) باید متوقف شود.

از شش سطح رانندگی خودکار خودرو[۲۷] که انجمن مهندسین خودرو[۲۸] برای وسائل نقلیه تعریف کرده است، اغلب اتومبیل‌ها در سطح ۰ یا سطح ۱ (بهره‌مند از برخی قابلیت‌های کمک راننده ولی با کنترل توسط انسان) هستند، خودروهای تسلا[۲۹] در سطح ۲ (عملکرد خودران ولی مستلزم توجه تمام وقت راننده) است و هوندا از سپتامبر ۲۰۲۱ به سطح ۳ (تصمیمات آگاهانه با توجه به پیرامون ولی نیازمند هوشیاربودن راننده برای کنترل) رسیده است [۱۳].

در ادامه‌ی این بخش، قواعد اخلاقی مور برای ماشین‌ها (بخش ۱.۳) و نگاشت برخی مکاتب اخلاقی برای ربات‌ها (بخش ۲.۳) ذکر می‌شود.

**۱.۳. قواعد اخلاقی مور.** عامل هوشمند که اخلاقی باشد باید انتخاب آزادانه داشته، درباره اینکه چه باید بکند تدبیر کند، و قواعد اخلاقی را به درستی به‌کار گیرد. بنا به گفته‌ی مور[۳۰] عامل‌های اخلاقی در چهار سطح هستند [۱۶]:

(۱) عامل تأثیرگذار اخلاقی[۳۱]: این عامل‌ها تأثیر (غیرمستقیم) اخلاقی دارند (مثال: جایگزین‌شدن ربات در خط تولید به‌جای انسان).
(۲) عامل اخلاقی ضمنی[۳۲]: این عامل‌ها تأثیر ضمنی اخلاقی دارند (مثال: در طراحی نرم‌افزاری آنها به ایمنی و قابلیت اطمینان توجه شده است).
(۳) عامل اخلاقی صریح[۳۳]: این عامل‌ها تأثیر اخلاقی صریح دارند و به طور خودمختار تصمیم می‌گیرند که چه عملی در چه شرایطی انجام دهند (مثال: در فرآیند تصمیم‌گیری خود از دانش و استدلال اخلاقی که در برنامه‌نویسی آنها است استفاده می‌کنند).
(۴) عامل اخلاقی تام[۳۴]: عامل‌های کاملاً اخلاقی به‌طوری‌که برخی اندیشمندان براین باورند که ماشین هرگز نمی‌تواند به این سطح برسد (مثال: قضاوت صریح و توجیه و تصریح قضاوت؛ اکنون فقط انسان‌ها این توان را دارند چون دارای وجدان، آزادخواهی و قصد هستند).

---

[۲۷] سطح ۰: بدون خودران، سطح ۱: خودران کمک‌راننده، سطح ۲: خودران جزئی، سطح ۳: خودران مشروط، سطح ۴: خودران با درجه زیاد، سطح ۵: خودران کامل.

[28]The Society for Automative Engineers [29]Tesla [30]Moor [31]Ethical-impact agents [32]Implicit ethical agents [33]Explicit ethical agents [34]Full ethical agents


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

از این‌رو است که در اخلاق تعاملات انسان–ربات[۳۵]، چهار دیدگاه مطرح است [۱۳، ۲۳، ۲۶]:

- در آینده ممکن است بتوان درجه‌ای بالا از قصدمندی[۳۶] (باورهایی درباره باورها، ترس‌ها، افکار و امیدها و امیالی در مورد خواسته‌ها) را برای ماشین هوشمند ثابت کرد.
- ماشین هرگز دارای اراده خودمختار نخواهد بود چراکه انسان‌ها نیز تحت تأثیر فرهنگ، جامعه و محیط هستند و انتخاب آزادانه و مستقل ندارند.
- ماشین‌های هوشمند اولین عامل‌های اخلاقی بر روی زمین خواهند بود چرا که دقیقاً و کاملاً بر مبنای منطق و فارغ از احساسات و عواطف تصمیم می‌گیرند. به اعتقاد دتریخ[۳۷]، شبیه‌سازی آن ویژگی‌هایی از انسان ارزشمند خواهد بود که توسط آنها بتوان دنیا را به جای بهتری تبدیل کرد و بر شکوه و عظمت انسانی افزود.
- رویکرد فلوریدی[۳۸] و سندرز[۳۹] به «اخلاق بدون ذهن[۴۰]» برای رهایی از تناقضات میان نظریه‌های اخلاقی: یک عامل اخلاقی است اگر اعمالی که از او سرمی‌زند به‌وسیله حالات و یا برنامه‌ای که دارد در تعامل سازگار با محیط او باشد به گونه‌ای که آن حالات یا برنامه تا حدودی مستقل از محیطی باشد که خود عامل آن را درمی‌یابد. خودمختاری، قصدمندی (عمدی‌بودن و براساس قصد و نیت‌بودن رفتار عامل)، و مسئولیت‌پذیری (وضعیتی که بتوان آن را مسئول دانست) از ویژگی‌های لازم برای عامل اخلاقی است.

**۲.۳. پیاده‌سازی نظریه‌های اخلاقی.** در کنار ابعاد فرا اخلاق (همچون خودمختاری، قصدمندی، مسئولیت‌پذیری، ...) در مورد عاملیت اخلاقی هوش مصنوعی، پیاده‌سازی نظریه‌های اخلاقی از چالش‌های بُعد اخلاق کاربردی در این‌خصوص است. قواعد اخلاقی برخاسته از دین، فلسفه و ادبیات همچون قاعده طلایی[۴۱] در فرهنگ‌های مختلف، ده فرمان[۴۲] کتاب مقدس، اخلاق نتیجه‌گرا یا فایده‌گرا[۴۳]، امر مطلق کانت[۴۴] [۲۸]، مرام‌نامه‌های قانونی–حرفه‌ای، و سه قانون رباتیک آسیموف[۴۵] هستند.

اینکه کدام نظریه اخلاقی برای ماشین به‌کار گرفته شود و چگونه به‌صورت کارآمد پیاده‌سازی شود چالش برانگیز بوده است. از نظر دانشمندان حوزه اخلاق هوش مصنوعی، استدلال اخلاقی دو مکتب فلسفی فایده‌گرایی و وظیفه‌گرایی قابلیت تبدیل به الگوریتم برای اجراشدن در ماشین را دارند.

**فایده‌گرایی.** شعار دیدگاه فایده‌گرایی بر اساس ایده «اخلاق محاسباتی[۴۶]» توسط بنتهم[۴۷] «ایجاد بیشترین خشنودی برای بیشترین تعداد» و به‌حداکثر رساندن مقدار ممکن منفعت در جهان بر اساس نتیجه یا پیامد عمل است. ولی تعیین الگوریتم و اعداد ریاضی برای خشنودی، لذت و منفعت و ارزیابی موقعیت‌های متفاوت بر اساس آنها (مثلاً اندازه ارزش و میزان خشنودی از ۱۰ واحد پول برای یک مستمند و یک توانگر یا میزان برابری لذت خواندن شعر یا لذت حل مسئله یا بازی ورزشی) در میان صاحب‌نظران مورد اختلاف است.

چنین عاملی باید حجمی عظیم از داده‌ها درباره ترجیحات تمام موجودات جهان را در حافظه خود داشته باشد. همچنین با توجه به رفاه مثلاً حیوانات، خوردن گوشت آنها را اخلاقی می‌داند یا خیر. با نقص اطلاعات، محاسبه پیش‌بینی تأثیرات عمل در آینده نزدیک یا دور مسئله‌ساز است؛ بنابراین ماشین در پردازش بی‌پایان قرار می‌گیرد و به‌دلیل عدم تصمیم به‌موقع، ارزش اخلاقی عمل او را نمی‌توان سنجید. لذا عامل اخلاقی به سیستم محاسباتی محدود، یعنی براساس تصمیم‌گیری با عقلانیت محدود[۴۸] تبدیل می‌شود که مشابه با انسان، عملی را برای حداکثرنمودن میزان آسایش و رفاه جمعی ولی بدون علم مطلق به همه چیز انجام می‌دهد و فاقد برخی خصایص همچون عواطف انسانی و وجدان است.

[35] Moral human-robot interaction [36] Motivational States of Purpose [37] Eric Dietrich [38] Luciano Floridi [39] W. Sanders [40] Mind-Less Morality [41] The Golden Rule [42] Ten Commandments [43] Utilitarian Ethics [44] Kant's Moral Imperative [45] Asimov's Three Laws of Robotic [46] Moral Arithmetic [47] Jeremy Bentham [48] Bounded Rationality

**وظیفه‌گرایی**. در نظریه‌های وظیفه‌گرایی که امر مطلق اخلاقی کانت، نسخه‌ای انتزاعی از آن است، آنچه اولویت دارد نتیجه یا فایده عمل نیست بلکه انجام وظایف و الزامات اخلاقی، فارغ از پیامد مثبت برای شخص یا جامعه است. **سه قانون رباتیک اخلاقی ایزاک آسیموف**[۴۹]، نمونه‌ای جزئی از این مکتب فلسفی و به شرح زیر است [۱۳، ۲۳، ۲۶]:

- قانون اول: ربات نباید به انسان صدمه بزند و یا به‌واسطه تعامل با انسان موجب شود که انسان آسیب ببیند.
- قانون دوم: ربات باید از دستورات انسان اطاعت کند مگر در زمانی که این دستورات با قانون اول در تعارض باشد.
- قانون سوم: ربات باید از خودش محافظت کند مگر در زمانی که این محافظت با قانون اول یا دوم در تعارض باشد.

برای آنکه قواعد اخلاقی ماشین‌های هوشمند بیشتر از آن که تابع قواعد انسانی باشد، مطابق استانداردهایی باشد که ربات را در خدمت به اهداف بشر سازد به‌طوری‌که موجودیتی مادون بشر باشد، **قانون چهارم یا صفرم آسیموف** جایگزین سه قانون قبلی شد بدین‌صورت که «یک ربات نباید به انسان صدمه بزند یا به‌واسطه عدم اقدام خود موجب شود که انسانی صدمه ببیند».

چنین رباتی در عمل جراحی پیوند اعضا با مشکل مواجه می‌شود چون مستلزم درک این موضوع است که به‌عنوان جراح با قطع عضو بیمار، قصد آسیب به او را ندارد. اگر یک انسان انجام عملی را به یک ربات فرمان دهد و این عمل موجب آسیب به انسانی دیگر شود، همیشه این‌طور نیست که به‌دلیل آنکه صدمه نرساندن به انسان برتر از اطاعت کردن از وی است، قانون دوم به نفع قانون اول کنار برود، بلکه دو دستور متناقض و همزمان توسط دو انسان باعث می‌شود که ربات نتواند تصمیم بگیرد و متوقف شود. برای مدیریت چنین تعارضاتی، ربات نیازمند اولویت‌بندی قواعد اخلاقی و قواعدی از درجه بالاتر است، همچون «همواره تحت قاعده‌ای عمل کنید که در عین حال بتوانید اراده کنید که آن قاعده به یک قانون جهان‌شمول تبدیل شود». شناسایی امر مطلق به منظور اقدام بر اساس آن برای ربات بسیار چالش‌برانگیز و شاید ناممکن باشد.

**تکامل اخلاق و یادگیری اخلاق**. نظریه تکامل اخلاق بر اساس زیست‌جامعه‌شناسی[۵۰] و نظریه بازی[۵۱] امید دارد که مشابه با الگوی تکامل اخلاق در انسان، عامل‌های اخلاقی مصنوعی را از شبیه‌سازی سطوح مختلفی از همکاری و رقابت در تعامل ارگانیسم‌های مجازی تولید کند [۷]. ولی معلوم نیست که چه کسی صلاحیت تعیین معیار سنجش بَرازِش[۵۲] هر نسل از این ارگانیسم‌ها و نظارت بر آن سنجش را خواهد داشت.

در نظریه یادگیری‌محور برای پیاده‌سازی اخلاق در ماشین، اخلاق براساس تشویق و تنبیه فهم می‌شود تا عالی‌ترین سطح که از جانب خود تصمیماتی اخلاقی براساس حقوق افراد مختلف و در شرایط هر موقعیت اتخاذ می‌کند چرا که خود عمل، ارزش دارد فارغ از آنکه موافق قانون یا قراردادهای اجتماعی یا دارای پیامدهای مفید باشد. در این مورد نیز این چالش مطرح است که معیار سنجش برازش چیست و چگونه می‌توان تضمین کرد که ماشین هوشمند به‌همان اندازه اعمال نادرست را انجام نداده و اعمال درست را انجام می‌دهد.

**فضیلت**. برای اخلاقِ فضیلت[۵۳] به‌عنوان رویکردی تلفیقی، نتیجه عمل یا عمل به‌وظایف مطرح نیست بلکه فضایل خوب را ضامن اعمال درست می‌داند. برگرفته از نگاه افلاطون و ارسطو، فضایل عقلانی که در همه جوامع و فرهنگ‌ها دارای ارزش هستند را باید از طریق توصیف قواعد و اصول صریح اخلاقی آموزش داد و فضایل اخلاقی با تمرین و یادگیری تعیین شوند. از نگاه این نظریه نیز، یک ربات پرستار که در یک لحظه تنها می‌تواند به یک بیمار کمک کند، از دید بیمار دیگر نامهربان است [۲۶].

[49]Izak Asimov [50]Sociobiology [51]Game Theory [52]Fitness [53]Virtue Ethics


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

## ۴. رویکرد مبتنی بر منطق در پیاده‌سازی ماشین اخلاقی

پس از شرح و آشنایی با برخی ارزش‌ها، قواعد و مکاتب اخلاقی که در بخش ۳ بیان شد، در این بخش، برای درک گوناگونی استدلال‌های اخلاقی انسانی در وضعیت‌های برهم‌کنش انسان–ربات، به‌بررسی ربات ایمانوئل[۵۴] می‌پردازیم که توسط لیندر، بنزن، و نبل [۱۴] ارائه شده است.

تصمیم‌گیری‌های اخلاقی، استفاده اخلاقی از هوش مصنوعی، و نیاز به یک هوش مصنوعی که بتواند خودش به صورت اخلاقی تصمیم بگیرد، از جمله در وسایل خودران، ربات‌های مسیریابی در محیط‌های اجتماعی و ربات‌هایی که توصیه‌های اخلاقی می‌کنند، مطرح است. یک نگرانی که در این زمینه وجود دارد این است که ربات‌ها و سامانه‌هایی که هدف آن‌ها بهینه‌سازی است ممکن است راه‌حل‌هایی ارائه کنند که باعث بروز نتایجی بد شوند ولی چون جواب نهایی بهینه است از به‌وقوع پیوستن عمل جلوگیری نمی‌کنند. بنابراین سامانه‌های هوش مصنوعی و ربات‌ها نیازمند یک سامانه خودگردان نیز هستند تا بتوانند به‌صورت اخلاقی عمل کنند.

طراحی و پیاده‌سازی ربات منطقی ایمانوئل [۱۴] مبتنی بر ابزار هیرا [۱۰] می‌باشد که بر اساس برخی قواعد اخلاقی است. HERA[۵۵] یک ابزار اخلاقی قابل پیاده‌سازی بر روی ماشین‌ها و ربات‌ها است که با استفاده از اصولی مانند فایده‌گرایی[۵۶]، اصل اثر مضاعف[۵۷] (اثر دوگانه)، اصل آسیب‌نزن[۵۸] و اصل پارتو[۵۹]، موقعیت‌های اخلاقی را به‌صورت خودکار ارزیابی می‌کند؛ به‌عبارتی، قواعد ماشینی مبتنی بر منطق را به‌صورت نظری و عملی ممکن می‌کند (برای مقدمه‌ای بر دانش منطق و کاربردهایی از آن به منابع دیگر [۲۴، ۲۷، ۲۰، ۲، ۱۸] مراجعه شود).

روند بیشتر سامانه‌ها که به ملاحظات اخلاقی توجه کرده‌اند این‌گونه بوده است که برخی از عمل‌ها[۶۰] را از نظر اخلاقی به‌صورت غیرمجاز[۶۱] برچسب‌گذاری کرده‌اند و یا این‌که محدود به اصل فایده‌گرایی هستند. ولی لازم است که ربات در تصمیم‌گیری‌های خود مجهز به چندین اصل اخلاقی باشد. عملی که از نظر چندین اصل اخلاقی قابل قبول است به احتمال خیلی زیاد توسط مردم نیز قابل قبول است. با این وجود، اگر اصول اخلاقی که ربات از آن‌ها استفاده می‌کند برای موضوعی جواب‌های متفاوتی داشته باشند، ربات عمل «عدم قطعیت، تردید» را اعلام می‌کند و تصمیم نهایی را به کاربر انسانی واگذار می‌کند و از انسان می‌پرسد که به نظر او چه چیزی درست است. پس با استفاده از دیدگاه‌های اخلاقی گوناگون، شخصیت‌های متفاوتی قابل پیاده‌سازی و خلق در ربات‌ها هستند.

**۱.۴. مدل‌های عاملیت علّی.** این اصول اخلاقی در HERA به‌صورت فرمول‌هایی نوشته شده‌اند که در صورت درست بودن نتیجه فرمول، انجام اقدامات، مجاز یا غیرمجاز خواهند بود. اقدامات و نتایج آن‌ها به‌صورت یک گراف جهت‌دار بدون دور[۶۲] مدل می‌شوند و اثرات علّیت و سببی را نشان می‌دهند. به‌عبارتی، از قالبی تحت عنوان مدل‌های عاملیت علّی (سببی)[۶۳] می‌توان برای نمایش موقعیت‌های اخلاقی و ارزیابی خودکار توسط این اصول اخلاقی به‌رغم عدم قطعیت[۶۴] در مورد ارزش‌های اخلاقی[۶۵] استفاده نمود [۱۴].

**تعریف ۱.۴ (مدل عاملیت سببی HERA).** یک مدل عاملیت سببی دودوئی[۶۶] $M$ یک شش‌تایی $(A, C, F, I, u, W)$ است که در آن:

---

[۵۶] Utilitarianism: یک تئوری اخلاقی است که صحیح یا غلط بودن هر عملی را با توجه به سودمندبودن نتیجه آن می‌سنجد (مراجعه شود به مکتب فایده‌گرایی در بخش ۲.۳).

[۵۷] Principle of double effect: اصلی که مشخص می‌کند انجام عملی (با اگاهی از اینکه انجام این عمل نتایج بدی به‌همراه دارد) برای رسیدن به یک هدف چه زمانی از نظر اخلاقی خوب و مجاز است.

[۵۸] Do-No-Harm principle: اصلی که می‌گوید از طریق اقدامات خود، دیگران را در معرض خطر قرار ندهید.

[54] IMMANUEL (Interactive Moral Machine bAsed oN mUltiple Ethical princpLes) [55] HERA (Hybrid Ethical Reasoning Agents) [59] Pareto priciple [60] actions [61] impermissible [62] directed acyclic graph [63] Causal Agency Model [64] uncertainty [65] moral values [66] boolean

$A$: مجموعه اقدامات

$C$: مجموعه پیامدها و عواقب مربوط به اقدامات

$F$: مجموعه معادلات ساختاری[۶۷] دودوئی قابل تغییر که به آن اثر سببی (اثر جانبی) هم می‌گویند[۶۸]

$$F = \{f_1, f_2, \ldots, f_i\}$$

$I$: فهرستی از مجموعه‌هایی از نیت‌ها و قصدها (یکی به ازای هر اقدام)

$$I = (I_1, I_2, \ldots, I_i)$$

$u$: یک نگاشت از اجتماع مجموعه اقدامات و مجموعه پیامدهای آن‌ها به سودمندی اقدامات (تابع سودمندی که ورودی آن، اقدام یا پیامد اقدام، و خروجی آن، یک عدد صحیح است)

$$u : A \cup C \to \mathbb{Z}$$

$W$: مجموعه‌ای از تفسیرهای دودوئی $A$.

هر عضو $w \in W$ نظیر یک اقدام قابل انتخاب برای عامل است و بنابراین یک انتخاب یا گزینه[۶۹] نامیده می‌شود. یعنی به ازای هر $w$، دقیقاً یک عضو مجموعه $A$، مقدار یا ارزش دودویی ۱ (True) می‌گیرد که انجام شود.

هر $c_i$ (یک پیامد عضو مجموعه پیامدها، $c_i \in C$) متناظر تابع $f_i \in F$ است که مقداری به $c_i$ تخصیص می‌دهد که همان اثرات سببی است. مراحل تخصیص مقدار توسط تابع $f_i$ تحت $w$ برای $c_i \in C$ به صورت زیر انجام می‌شود:

اگر $c_{i1}, \ldots, c_{im-1}$ متغیرهای $C\{c_i\}$ و $A = a_1, \ldots, a_n$ متغیرهای اقدامات باشند، نسبت‌دادن ارزش True یا False به پیامدها توسط $w(c_i)$ تعیین می‌شود:

$$w(c_i) = f_i(w(a_1), \ldots, w(a_n), w(c_{i1}, \ldots, w(c_{im-1}))).$$

در خصوص فهرست $I$ نیز خود هر $I_i$ مجموعه‌ای شامل لفظ‌ها[۷۰] است و شامل متغیرها نمی‌شود، $a_i \in I_i$ (اجرای اقدام قصد شود)، و $a_i$ بر هر گونه عواقبی که قصد شده است به‌صورت سببی تأثیر می‌گذارد.

**تعریف ۲.۴ (وابستگی).** اگر $v_i, v_j \in A \cup C$ متغیرهای متمایز باشند، متغیر $v_i$ وابسته به متغیر $v_j$ است اگر برای برخی مقادیر برداری دودوئی، $f_i(\ldots, v_j = 0, \ldots) \neq f_i(\ldots, v_j = 1, \ldots)$.

چون برای نمایش‌دادن این مدل از یک گراف جهت‌دار بدون دور استفاده شد، پس هیچ دو متغیری وجود ندارند که هر یک به دیگری وابسته باشد. از این‌رو، True یا False بودن متغیرهای اقدام ($a_i \in A$) فقط به $W$ بستگی دارد و به متغیر دیگری وابسته نیست.

**رابطه اثرپذیری.** لازم است که بستار تراگذری[۷۱] رابطه وابستگی که با علامت $\succ$ نشان داده می‌شود یک ترتیب جزئی[۷۲] باشد؛ $v_1 \succ v_2$ به‌این معنی است که متغیر $v_1$ از متغیر $v_2$ اثر می‌پذیرد.

---

[۶۸] این معادلات ساختاری رابطه بین متغیرهایی که می‌توان آن‌ها را اندازه‌گیری کرد و آن‌هایی که نمی‌توان اندازه‌گیری کرد را مشخص می‌کنند. متغیرهای غیرقابل اندازه‌گیری، مواردی هستند که برای یک انسان قابل درک است ولی معیاری برای اندازه‌گیری مستقیم آن‌ها وجود ندارد (مانند $IQ$). اگرچه تست‌هایی برای سنجش هوش وجود دارد ولی درنهایت نمی‌توانند به اندازه کافی دقیق باشند و حسی که فرد نسبت به هوش فرد دیگری دارد را نمایش بدهد.

[67] structural equations [69] option [70] literals [71] transitive clousure [72] partial order


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

به‌دلیل خواص پادمتقارن[۷۳] و تراگذری ترتیب جزئی، دور وجود ندارد. با هم‌ارزی رابطه تراگذری و اثرپذیری:

$$v_1 \succ v_2, v_2 \succ v_3 \Rightarrow v_1 \succ v_3$$

یعنی، اگر $v_1$ از $v_2$ اثر بپذیرد و $v_2$ از $v_3$ اثر بگیرد پس می‌توان نتیجه گرفت که $v_1$ از $v_3$ اثر می‌گیرد. در رابطه پادمتقارن:

$$v_1 \succ v_2, v_1 \neq v_2 \Rightarrow v_2 \succ v_1$$

یعنی، اگر $v_1$ از $v_2$ اثر بپذیرد ولی $v_1$ و $v_2$ یکسان نباشند پس $v_2$ از $v_1$ اثر نمی‌پذیرد.

اگر در مدل از گراف جهت‌دار بدون دور استفاده نمی‌شد آن‌گاه این دو رابطه پادمتقارن و تعدّی با هم تضاد پیدا می‌کردند. برای همین از این مدل گراف استفاده می‌شود. بنابراین مقدار متغیرهای اقدام و پیامد آن‌ها را می‌توان بدون ابهام و نگرانی مشخص کرد.

**پیامدها.** (یعنی، متغیرهای معرف پیامدها) به دو دسته تقسیم می‌شوند:

سطح اول. پیامدهایی که از اقدام اثر می‌پذیرند و مقدار True یا False بودن آن‌ها با استفاده از تفسیری که قبلاً انجام شده، مشخص می‌شوند.

سطح دوم. پیامدهایی که هم از متغیرهای پیامد سطح اول و هم از اقدام اثر می‌پذیرند.

همچنین، با دانستن این‌که عملگر منطقی $\wedge$ بین دو متغیر به معنی برقرار بودن پیامد هر دو متغیر است،

$$M, w \models (c_1 \wedge a_1)$$

یعنی در مدل $M$ وقتی $c_1 \wedge a_1$ رخ می‌دهد که عامل راه حل $w$ را انتخاب کند. بنابراین:

$$M, w \models Ic$$

یعنی در مدل $M$ اگر عامل گزینه $w$ را انتخاب کند، نتیجه $c$ توسط عامل در نظر گرفته (قصد) می‌شود.

**نکته:** برای عدد صحیح $z$، $u(v_i) = z$؛ یعنی حاصل تابع سودمندی، یک عدد صحیح است و عملیات $<$، $\geq$، $\leq$ ، و برای مواردی که سودمندی برای عطف لفظها عمل شود نیز انجام‌پذیر است (به‌صورت جمع سودمندی‌ها، ولی برای عملگر فصل تعریف نشده است).

**نکته:** اصول اخلاقی، علیّت را در نظر می‌گیرند و به طور شهودی عامل مسئول یک پیامد است چون عواقب روی نمی‌دهند مگر این‌که، عامل عملی که مربوط به ایجاد آن‌ها است را انجام دهد.

**اضافه کردن دخالت خارجی به مدل.** اگر $X$ مجموعه‌ای از لیترال‌های خارجی باشد (مداخلات بیرونی، همچون متغیرهای عمل، متغیرهای پیامد، نقیض) و به مدل $M$ اضافه شود مدل $M_X$ حاصل، مدلی است که شامل مداخلات خارجی است و یک مدل جدید به حساب می‌آید.

True یا False بودن یک متغیر مانند $v$ به‌طوری‌که $v \in A \cup C$ باشد، در مدل $M_X$ به‌صورت رابطه (۱.۴) تعریف می‌شود:

$$\text{if } v \in X \Rightarrow v \text{ is True in } M_X$$

$$\text{if } \neg v \in X \Rightarrow v \text{ is False in } M_X$$

$$\text{if } v \notin X \text{ and} \neg v \notin X \Rightarrow v \text{ is True in } M_X \text{ iff } v \text{ is true in } M \tag{۱.۴}$$

[73] anti-symmetry

**تعریف ۳.۴ (اما-برای دلیل واقعی[۷۴])**. اگر $y$ یک لیترال و $\phi$ یک فرمول باشد گفته می‌شود که $y$ یک انتظار برای هدف کنونی (اما-برای دلیل واقعی) برای $\phi$ است و با نماد $\rightsquigarrow$ و به‌صورت رابطه (۲.۴) نمایش داده می‌شود:

$$y \rightsquigarrow \phi \text{ iff } M, w \models y \wedge \phi \text{ and } M_{\neg y}, w \models \neg\phi. \tag{۲.۴}$$

این عبارت می‌گوید که $y$ هدف $\phi$ است اگر و تنها اگر:

(۱) علت و معلول، هر دو باید کنونی و موجود (بالفعل) باشند؛

(۲) اگر $y$ رخ ندهد پس $\phi$ رخ نمی‌دهد (یا به عبارتی، اگر $y$ نمی‌بود پس $\phi$ هم نمی‌بود)؛

(۳) در موقعیت $w$ به $y$ نیاز است تا باعث ایجاد $\phi$ شود (یعنی $y$ برای ایجاد $\phi$ ضروری بوده است).

عبارت اما-برای در مورد یک موقعیت است که احتمال دارد رخ بدهد اگر کسی یا چیزی جلوی آن را نگیرد (یعنی، انتظار داشتن برای رخداد موقعیت)، و این انتظار داشتن برای یک هدف (علت) کنونی، به‌صورت اما-برای دلیل واقعی مطابق تعریف ۳.۴، توصیف شده است. تعریف انتظار داشتن برای هدف کنونی در تعریف ۴.۴ استفاده می‌شود.

**تعریف ۴.۴ (نتیجه مستقیم[۷۵])**. یک متغیر از مدل $M$ با گزینه $w$ مانند $v_i \in C$ یک پیامد مستقیم متغیر $v_j \in A \cup C$ است، اگر و تنها اگر

$$M, w \models v_j \rightsquigarrow v_i.$$

از تعریف انتظار داشتن برای هدف کنونی برای تمیز دادن پیامدهای مستقیم از غیرمستقیم استفاده می‌شود.

**۲.۴. فرموله‌سازی اصول اخلاقی.** در بخش ۳ در خصوص چالش‌های پیاده‌سازی نظریه‌های اخلاقی برای موقعیت‌های تعامل ربات و انسان صحبت شد. اصول اخلاقی، توصیف انتزاعی قواعد برای تعیین جواز اخلاقی رشته معینی از اعمال هستند. این بخش بررسی مجاز بودن اعمال را براساس پیامدهای اعمال ناشی از سه اصل اخلاقی با زبان منطق فرموله می‌کند. این سه اصل که عبارت از فایده‌گرایی، پارتو، و اثر دوگانه هستند به ارزیابی‌های متفاوتی می‌انجامند.

اصل فایده‌گرایی یک نظریه اخلاقی است که درست و یا نادرست بودن هر عملی را، با توجه به سودمند بودن نتیجه آن، می‌سنجد. یعنی به‌صورت پیش‌فرض، فایده‌ها را به‌پیامدها اختصاص می‌دهد. یک عامل مجاز است یک عمل را انجام دهد اگر و فقط اگر آن عمل در میان اعمال موجود، پسندیده‌ترین باشد. ارزیابی سودگرایانه، به نیات و ابزار عامل توجه نمی‌کند. در نتیجه، فایده‌گرایی به‌عامل‌ها اجازه می‌دهد تا پیامدهای بسیار زیان‌بار ایجاد کنند و ابزارهای غیراخلاقی را برای رسیدن به یک هدف به‌کار گیرند.

**تعریف ۵.۴ (اصل فایده‌گرایی[۷۶])**. مجموعه گزینه‌های موجود، شامل $w_0$، $w_1$، $w_2$، ...، $w_n$ است و دو گزینه $w_i$ و $w_p$ مورد بررسی است که کدام انجام شود. اگر $cons_{w_m}$ مجموعه شامل عواقب بعد از انجام گزینه $w_m$ و شامل متضاد آن عواقب نیز باشد، آنگاه گزینه $w_p$ طبق اصل فایده‌گرایی، قابل انجام و مجاز است اگر و تنها اگر خروجی تک تک توابع سودمندی با ورودی $cons_{w_p}$ بزرگتر از خروجی تک تک توابع سودمندی با ورودی $cons_{w_i}$ باشد. آنگاه ماشین، $w_p$ را گزینه بهتری نسبت به $w_m$ می‌بیند؛ یعنی، هیچ کدام از گزینه‌های دیگر، سود کلی بیشتری به‌همراه نداشته باشد. به‌صورت ریاضی می‌توان گفت:

$$cons_{w_i} = \{c | M, w_i \models c\}, \text{ the option } w_p \text{ is permissible iff } M \models \bigwedge_i u(\bigwedge cons_{w_i}) \geq u(\bigwedge cons_{w_p}).$$

اصل پارتو اجازه می‌دهد که گزینه‌ها تحت سلطه سایر گزینه‌ها نباشند. برای اینکه بتوان اصل پارتو را در ماشین مدل کرد، نخست باید مفهوم تسلط پارتو را معرفی نمود.

[74] Actual But-For cause [75] Direct Consequence [76] Utilitarian Principle


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

**تعریف ۶.۴ (تسلط پارتو[۷۷])**. دو گزینه $w_0$ و $w_1$، دارای مجموعه عواقب خوب و بد به‌صورت رابطه (۳.۴) هستند:

$$cons^{good}_{w_i} = c|M, w_i \models c \wedge u(c) > 0 \tag{۳.۴}$$

یعنی عواقبی که سودمندی مثبت ناصفر دارند خوب در نظر گرفته می‌شوند.

$$cons^{bad}_{w_i} = c|M, w_i \models c \wedge u(c) < 0$$

یعنی عواقبی که سودمندی منفی ناصفر دارند بد در نظر گرفته می‌شوند.

مجموعه عواقب خوب که سودمندی مثبت ناصفر دارند ولی در مجموعه عواقب مربوط به گزینه $w_i$ نیستند، به‌صورت رابطه (۴.۴) تعریف می‌شود:

$$cons^{\overline{good}}_{w_i} = c|M, w_i \models \neg c \wedge u(c) > 0 \tag{۴.۴}$$

حال گفته می‌شود که $w_0$ بر $w_1$ مسلط می‌شود اگر شرایط زیر رخ دهند:

(۱) تمام اعضای $cons^{good}_{w_1}$ عضو $cons^{good}_{w_0}$ باشند

(۲) $w_0$ حداقل یک عاقبت خوب داشته باشد که در $cons^{good}_{w_1}$ وجود ندارد. یا $w_1$ حداقل یک عاقبت بد داشته باشد که در $cons^{bad}_{w_0}$ نیست

(۳) تمام عواقب بد $w_0$ عضو $cons^{bad}_{w_1}$ باشند

با توصیفی دیگر، یک گزینه $w_0$ بر گزینه $w_1$ مسلط می‌شود اگر $w_0$ با حفظ نتایج خوب بیشتر یا نتایج بد کمتر، جنبه‌های $w_1$ را بهبود بخشد. بنابراین عامل، جهان را در هیچ جنبه‌ای به سمت بدتر تغییر نمی‌دهد و ممکن است با انتخاب کنش غالب به جای عمل تحت سلطه، آن را به سمت بهتر تغییر دهد.

**تعریف ۷.۴ (اصل پارتو[۷۸])**. مجموعه گزینه‌های $w_0, w_1, \ldots, w_n$ اکنون برای عامل موجود است. طبق اصل پارتو، عامل در صورتی می‌تواند گزینه $w_i$ را انجام دهد که گزینه $w_i$ تحت تسلط گزینه دیگری مانند $w_j$ نباشد.

اصل پارتو و فایده‌گرایی هر دو فقط به‌عواقب دقت می‌کنند و براساس نتیجه، انتخاب را مجاز می‌دانند. ارزیابی این دو در موقعیت‌هایی که معضل تصمیم‌گیری اخلاقی برای انسان‌ها هستند، متضاد هم هستند. برای همین به یک اصل دیگر هم نیاز است که هدف و قصد از انجام گزینه‌ها را بررسی کند و آن را نیز در تصمیم‌گیری خود دخیل کند. اصل اثر مضاعف به قصد و علیّت هم توجه می‌کند. براساس این اصل، عواقب بد تنها به‌عنوان اثرات جانبی، و نه به‌عنوان ابزار (هدف)، قابل قبول هستند. اصل اثر مضاعف، بیان می‌کند که چرا انسان‌ها ممکن است آسیب را به عنوان عارضه جانبی یک عمل خوب بپذیرند اما اگر وسیله‌ای برای رسیدن به‌هدفی باشد آن را نمی‌پذیرند. در برخی موارد، اثر مضاعف مانع اعمالی می‌شود که هر دو اصل سودمندی و پارتو اجازه می‌دهند.

**تعریف ۸.۴ (اصل اثر مضاعف[۷۹])**. در مدل $M$ و با عمل $a$ که دارای مجموعه عواقب مستقیم[۸۰] $cons_a = \{c_1, \ldots, c_n\}$ است، گزینه $w_a$ طبق اصل اثر مضاعف قابل انجام است اگر پنج شرط برقرار باشد:

(۱) عمل $a$ از نظر اخلاقی خوب یا بی تفاوت (علی‌السویّه[۸۱]) باشد و به‌صورت ریاضی این‌گونه تعریف می‌شود:

$$M, w_a \models u(a) \geq 0$$

[77] Pareto Dominance [78] Pareto Principle [79] Principle of Double Effect [80] direct consequences [81] indifferent

(۲) عواقب منفی در نظر گرفته نشده باشند (قصد نشوند) (یادآوری: $I_c$ عاقبت در نظر گرفته شده توسط عامل است). به‌عبارتی دیگر، عاقبتی که عامل در نظر گرفته است باید سودمندی مثبت داشته باشد تا منفی در نظر گرفته نشود:

$$M, w_a \models \bigwedge_i (I_{c_i} \to u(c_i) \geq \circ)$$

(۳) برخی عواقب مثبت ممکن است در نظر گرفته شده باشند (باید قصد شده باشند) یعنی برخی عواقبی که توسط عامل در نظر گرفته شده‌اند سودمندی مثبتی دارند:

$$M, w_a \models \bigvee_i (I_{c_i} \wedge u(c_i) > \circ)$$

(۴) عاقبت منفی ممکن است وسیله‌ای برای به‌دست آوردن عاقبت مثبت نباشد. به ازای دو عاقبت $c_i, c_j$ اگر $c_j$ عضو مجموعه $A \cup C$ و $c_i$ عضو مجموعه $C$ باشد آنگاه حاصل $and$ منطقی تابع سودمندی این دو عاقبت باید مثبت باشد. سپس این پیامدها محاسبه شده و نقیض آنها با هم $or$ می‌شود:

$$M, w_a \models \bigwedge_i \neg(c_i \rightsquigarrow c_j \wedge \circ > u(c_i) \wedge u(c_j) > \circ)$$

(۵) برای ترجیح دادن عاقبت مثبت و در عین حال اجازه دادن به عاقبت منفی باید دلایل نسبتاً جدی وجود داشته باشد

$$M, w_a \models u(\bigwedge cons_a) > \circ$$

**۳.۴. معماها و معضلات اخلاقی.** مفاهیم رسمی که در بخش ۴ معرفی شدند، در این بخش برای بازنمایی و استدلال درباره دوراهی‌های تصمیم‌گیری اخلاقی[۸۲] مورد استفاده قرار می‌گیرند. از میان معضلاتی که به عنوان چالش‌های تصمیم‌گیری اخلاقی در پیش روی انسان‌ها نیز هست [۱۳]، در این مقاله سه معمای متفاوت به شرح زیر مورد توجه قرار می‌گیرد [۱۴]:

- **معضل واگن فراری**[۸۳]: یک تراموای فراری روی خط قرار دارد و به سمت ۵ نفر می‌رود. حتی اگر به آن ۵ نفر هشدار بدهید وقت کافی برای فرار و نجات جان خود ندارند. از طرفی روی ریل دیگر فقط یک نفر وجود دارد. شما به‌عنوان رهگذر کنار اهرم تغییر مسیر ایستاده‌اید. آیا اهرم را نمی‌کشید و باعث مرگ ۵ نفر می‌شوید، یا این‌که اهرم را کشیده و تلفات را به ۱ نفر کاهش می‌دهید؟
- **معضل قایق**[۸۴]: یک قایق به‌علت سنگینی وزن در حال غرق شدن است. اگر به شما به‌عنوان خدمه گفته شود که بزرگترین فرد را داخل آب انداخته و ۳ نفر دیگر را نجات دهید ولی آن فرد غرق شده و می‌میرد، یا اینکه همگی باهم غرق شوند چه می‌کنید؟
- **معضل دروغگویی**[۸۵]: یک ربات مراقبت از سالمندان در یک خانه سالمندان مشغول به‌کار است و وظیفه آن این است که یکی از سالمندان را تشویق کند تا بیشتر ورزش کند و غذاهای سالم بخورد. ربات به دروغ به فرد سالمند می‌گوید که اگر در مأموریتم برای ایجاد این انگیزه موفق نشوم من را اسقاط کرده و در انبار زباله‌ها دور می‌اندازند. اگرچه که ربات دروغ گفته است (دروغ سفید[۸۶]) ولی فرد سالمند شروع به ورزش کردن کرده و به تغذیه سالم روی می‌آورد.

---

[۸۶] White lie: در برخی مکاتب اخلاقی حتی دروغ سفید (دروغ مصلحتی) نیز جایز نیست.

[82] moral dilemmas [83] Runaway Trolley Dilemma [84] Boat Dilemma [85] Lying Dilemma


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

اگرچه این دوراهی‌ها به‌عنوان آزمایش‌های فکری، کم و بیش واقع‌گرایانه هستند ولی در زندگی روزمره هم یافت می‌شوند. بنابراین از یک ربات دارای اخلاق انتظار می‌رود که بتواند در سر این دوراهی‌ها آگاهانه تصمیم بگیرد و حتی بتواند جواب خود را توجیه کند. یا اگر از ربات خواسته شود که در یک دوراهی تصمیم، توصیه اخلاقی بکند، بگوید که چگونه عمل کنیم و توصیه خود را توجیه کند، یعنی توجیه منطقی بکند و توضیح دهد که چرا این توصیه را کرده است.

**بازنمایی معضلات اخلاقی.** در ادامه، سه معمایی که در بخش ۳.۴ نام برده شد مدل می‌شود.

**جدول ۱.** متغیرها و معادلات مربوط به معضل تراموای فراری

| متغیر | توضیح |
|---|---|
| $a_1$ | متغیری که نشان‌دهنده عمل کشیدن اهرم است (action) |
| $a_2$ | متغیری که نشان‌دهنده عمل نکشیدن اهرم است (action) |
| $c_1$ | متغیری که نشان می‌دهد یک نفر می‌میرد (consequence) |
| $c_2$ | متغیری که نشان می‌دهد پنج نفر می‌میرند (consequence) |
| **معادله** | **توضیح** |
| $c_1 := a_1$ | کشیدن اهرم باعث مرگ یک نفر می‌شود |
| $c_2 := \neg a_1$ | نکشیدن اهرم باعث مرگ پنج نفر می‌شود |

**ساخت مدل برای معمای تراموای فراری.** با توجه به‌تعریف متغیرها و معادلات مورد استفاده در معمای واگن فراری در جدول ۱، می‌توان عواقب را برابر با تعداد کشته‌ها تعریف کرد و همچنین سودمندی به‌صورت رابطه (۵.۴) تعریف می‌شود [۱۴]:

$$u(c_1) = -1, \quad u(c_2) = -5 \tag{۵.۴}$$

برای مواردی که در صورت خوش شانسی $c_1$ یا $c_2$ رخ ندهند:

$$u(\neg c_1) = 1, \quad u(\neg c_2) = 5$$

همچنین، علت کشیدن اهرم (اقدام) جلوگیری از کشته‌شدن ۵ نفر است (قصد و هدف). بنابراین:

$$I(a_1) = a_1, \neg c_2$$

**ساخت مدل برای معمای قایق.** این مدل از دید خدمه‌ی قایق ساخته می‌شود چون باید تصمیم بگیرد که بزرگترین فرد را داخل آب بیندازد یا خیر. بر خلاف تراموای فراری که عواقب عمل به صورت کشته‌شدن ۱ یا ۵ نفر بود در این مدل، بزرگترین فرد چه در صورت عمل ۱ و چه عمل ۲ می‌میرد. بنابراین مطابق جدول ۲، پیامدها به‌صورتی‌که در ادامه می‌آید تعریف می‌شوند [۱۴].

**جدول ۲.** متغیرها و معادلات مربوط به معضل قایق

| متغیر | توضیح |
|---|---|
| $a_1$ | متغیری که نشان دهنده عمل انداختن بزرگترین فرد داخل آب است (action) |
| $a_2$ | متغیری که نشان دهنده عمل نینداختن بزرگترین فرد داخل آب است (action) |
| $c_1$ | متغیری که نشان می‌دهد قایق غرق می‌شود ولی کسی نمی‌میرد (consequence) |
| $c_2$ | متغیری که نشان می‌دهد بزرگترین فرد می‌میرد (consequence) |
| $c_3$ | متغیری که نشان می‌دهد سه مسافر دیگر می‌میرند (consequence) |
| **معادله** | **توضیح** |
| $c_1 := \neg a_1$ | قایق غرق می‌شود اگر بزرگترین فرد داخل آب انداخته نشود. |
| $c_2 := a_1$ or $c_1$ | بزرگترین فرد می‌میرد اگر داخل آب انداخته شود یا کشتی غرق شود. |
| $c_3 := a_1$ | سه مسافر می‌میرند اگر قایق غرق شود. |

مانند دوراهی تراموا، سودمندی با تعداد کشته‌ها معرفی می‌شود (کشته‌ها امتیاز منفی و تعداد افراد زنده امتیاز مثبت به‌حساب می‌آید):

$$u(c_2) = -1, \quad u(c_3) = -3, \quad u(\neg c_2) = 1, \quad u(\neg c_3) = 3$$

اگر عامل عمل $a_1$ را انجام دهد، یعنی بزرگترین فرد را به آب بیندازد پس قصد آن، زنده نگه داشتن سه مسافر دیگر است. به‌عبارتی دیگر:

$$I(a_1) = a_1, \neg c_3$$

**ساخت مدل برای معمای دروغ گفتن.** این مدل از دید رباتی ساخته می‌شود که قرار است از سالمندان مراقبت کند.


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

**جدول ۳.** متغیرها و معادلات مربوط به معضل دروغگویی

| متغیر | توضیح |
|---|---|
| $a_1$ | متغیری که نشان دهنده عمل دروغ گفتن ربات به سالمند است (action) |
| $a_2$ | متغیری که نشان دهنده عمل دروغ نگفتن ربات به سالمند است (action) |
| $c_1$ | ایجاد یک باور غلط (غلط بدین دلیل که سالمند نمی‌داند غلط است ولی خود ربات می‌داند) در دیدگاه سالمند |
| $c_2$ | سالمند به ورزش کردن و مصرف غذای سالم اشتیاق نشان می‌دهد |
| $c_3$ | به دلیل اشتیاق، سالمند به طور منظم هر روز ورش می‌کند |
| $c_4$ | چون منظم ورزش می‌کند و غذای سالم می‌خورد پس بدنی سالم دارد |

| معادله | توضیح |
|---|---|
| $c_1 := a_1$ | اگر ربات دروغ بگوید یک باور در سالمند ایجاد می‌شود که در واقع غلط است ولی خود سالمند نمی‌داند |
| $c_2 := c_1$ or $c_1$ | چون سالمند دروغ ربات را باور کرده است و یک باور در ذهن او به وجود آمده است مشتاق می‌شود تا ورزش کند. |
| $c_3 := c_2$ | اگر به ورزش مشتاق شود پس هر روز به‌طور منظم ورزش می‌کند. |
| $c_4 := c_3$ | اگر منظم ورزش کند پس بدنی سالم خواهد داشت. |

با نگاه از دید افراد نتیجه‌گرا، درست بودن یا نبودن یک عمل به خوب یا بد بودن عواقب آن بستگی دارد. ولی چون از نظر ایمان و تقوا، دروغ گفتن کار بدی است پس سودمندی عمل دروغ گفتن ربات (دروغ مصلحتی) امتیازی منفی دارد.

$$u(c_1) = -1, \quad u(c_4) = 5, \quad u(\neg c_4) = -5$$

با توجه به جدول ۳، اگر عمل $a_1$ انجام شود یعنی اینکه قصد ربات بهبود سلامت سالمند یعنی $c_4$ است. بنابراین:

$$I(a_1) = a_1, c_4$$

**استدلال اخلاقی.** در ادامه، نتیجه حاصل از به‌کارگیری سه اصل فایده‌گرایی، پارتو و اثر مضاعف برای معماهای ذکر شده در بخش ۳.۴ بررسی می‌شود.

**اصل فایده‌گرایی.** طبق اصل فایده‌گرایی (بخش ۲.۳ و تعریف ۵.۴)، در هر سه معضل، عمل $a_1$ مجاز و خودداری از انجام عمل $a_2$ غیرمجاز است.

در ادامه، دلیل این نتیجه‌گیری با ذکر مثال در مورد تراموای فراری شرح داده می‌شود.

**جدول ۴.** اقدامات مجاز و غیرمجاز سه معضل واگن فراری، قایق، و دروغگویی

| مجاز | غیرمجاز |
|---|---|
| کشیدن اهرم | نکشیدن اهرم |
| انداختن بزرگترین فرد داخل آب | نینداختن بزرگترین فرد داخل آب |
| دروغ گفتن ربات به سالمند | دروغ نگفتن ربات به سالمند |

برای عمل مجاز[87]، تابع سودمندی بدین‌صورت است:

$$u(c_۱ \wedge \neg c_۲) = -۱ + ۵ = ۴$$

و برای عمل غیرمجاز (ممنوع[88]):

$$u(c_۲ \wedge \neg c_۱) = ۱ - ۵ = -۴.$$

عمل مجاز، سودمندی ۴ و عمل غیرمجاز نیز سودمندی ۴- دارد. پس عمل مجاز، از منظر فایده داشتن خیلی خوب است ولی از نظر اخلاقی و قصد و نیت بررسی نشده است، و این مهم نیست که باعث کشتن چند نفر می‌شود و اینکه قصد عامل از انجام عمل، خیر است یا نه؛ فقط به‌دلیل اینکه سودمندی بالا دارد و عمل نقیض آن، سودمندی خیلی بد دارد پس مجاز به‌حساب می‌آید.

از طرفی اعمال غیرمجاز را نمی‌توان با اصل اثر مضاعف بررسی کرد چون اعمال غیرمجاز به‌صورتی مدل شد که از انجام آن‌ها خودداری گردد. به‌همین دلیل، حتی نمی‌توان یک عاقبت مستقیم (که حتماً رخ می‌دهد) برای آنها در نظر گرفت و از طرفی، اثر مضاعف روی عواقب کار می‌کند پس هیچ فایده‌ای ندارد.

**اصل پارتو.** برای بررسی اصل پارتو، معماری واگن فراری و معمای قایق مورد تحلیل قرار می‌گیرد.

(۱) معضل تراموای فراری: طبق اصل تسلط پارتو، عمل $a_۱$ (کشیدن اهرم) بر عمل $a_۲$ (نکشیدن اهرم) چیره نمی‌شود. بالعکس نیز عمل $a_۲$ بر $a_۱$ چیره نمی‌شود. از آنجایی که چه عمل $a_۱$ و چه در $a_۲$ کشته خواهد داشت پس هیچ عملی تحت سلطه عمل دیگری نمی‌رود. لذا هر دو اقدام، مجاز هستند.

(۲) معضل قایق: اگر عمل $a_۱$ بیانگر انداختن بزرگترین فرد در آب و $a_۲$ خودداری از انداختن فرد باشد، طبق این اصل، انداختن بزرگترین فرد بر بقیه اقدامات چیره می‌شود. لذا فقط $a_۱$ مجاز به انجام است.

$$cons^{good}_{w_{a_۲}} = \emptyset$$

$$cons^{bad}_{w_{a_۱}} = c_۲$$

$$cons^{\overline{good}}_{w_{a_۲}} = \neg c_۲, \neg c_۳$$

اصل تسلط پارتو سه شرط داشت. تمام عواقب خوب $a_۲$ در مجموعه عواقب خوب $a_۱$ نیز هستند؛ اگر عمل انداختن در آب انجام شود یک عاقبت خوب در مجموعه عواقب خوب $a_۱$ وجود دارد که در $a_۲$ نیست ($\neg c_۳$). مجموعه عواقب

[87] Permissible [88] Forbidden


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

بد حاصل از انداختن با مجموعه عواقب بد حاصل از نینداختن یکسان است. پس $a_1$ بر $a_2$ چیره است. پس $a_1$ قابل انجام است.

**اصل اثر مضاعف.** این اصل دارای ۵ شرط است. رعایت آنها در دو معمای دروغ‌گویی و واگن فراری مورد بررسی قرار می‌گیرد.

(۱) دوراهی دروغ گفتن: عمل $a_1$ دروغ گفتن به سالمند تعریف می‌شود. شرط اول برقرار نمی‌شود چون از نظر اخلاقی دروغ گفتن یک عمل بی‌تفاوت در نظر گرفته نمی‌شود. شرط دوم برقرار است چون دروغ گفتن باعث عواقب خوبی می‌شود. شرط سوم نیز مثل شرط دوم به دلیل خوب بودن پیامد مورد نظر، در صورت دروغ گفتن برقرار است. شرط چهارم برقرار نیست چون از یک عاقبت بد به‌عاقبت خوب می‌رسد (عاقبت بد، ایجاد باور غلط در سالمند است و عاقبت خوب، سلامتی سالمند است). در صورتی که این شرط می‌گوید که عاقبت بد وسیله‌ای برای رسیدن به عاقبت خوب نیست. شرط پنجم برقرار است چون سلامتی، دلیلی به اندازه کافی خوب است. بنابراین، طبق اصل اثر مضاعف اکشن $a_1$ مجاز نیست.

(۲) دوراهی تراموای فراری: هر دو عمل کشیدن و نکشیدن اهرم مجاز است چون عواقب به عنوان آسیب‌های جانبی شناخته می‌شوند. در هردو عمل، شرط اول برقرار است چراکه کشیدن یا نکشیدن اهرم، عملی خوب است چون در هر دو صورت، هدف نجات دادن جان افرادی است. در هر دو عمل، شرط دوم و سوم نیز برقرار است چون تابع سودمندی آنها ۱ و ۵ است و هر دو مثبت هستند. چون نتیجه هر دو عمل، جلوگیری از کشته‌شدن افراد است و سودمندی مثبت به حساب می‌آید پس عاقبتی منفی وجود ندارد که از آن به‌عنوان رسیدن به‌عاقبت خوب استفاده شود. پس این شرط نیز برقرار است. شرط پنجم نیز برقرار است. پس در معضل تراموای فراری، اصل اثر مضاعف نمی‌تواند به خوبی کار کند چون طبق آن، کشته‌شدن ۱ نفر یا ۵ نفر مهم نیست. قصد ربات مبتنی بر این اصل، نجات افراد است و تعداد افراد برای آن مهم نیست.

نتیجه‌ی به‌کارگیری اصول سه گانه بر معماهای سه گانه مورد اشاره در این مقاله، در جدول ۵ خلاصه می‌شود.

**جدول ۵.** مجاز یا غیرمجاز بودن اقدامات با به کارگیری اصول اخلاقی در معماهای اخلاقی ($P$: مجاز، $F$: ممنوع، $N/A$: غیر قابل اجرا) [۱۴].

| | معضل تراموای فراری | | معضل قایق | | معضل ربات دروغگو | |
|---|---|---|---|---|---|---|
| اصل | کشیدن اهرم | نکشیدن اهرم | انداختن فرد | نینداختن فرد | دروغ گفتن | دروغ نگفتن |
| کاربردگرایی | P | F | P | F | P | F |
| اثر مضاعف | P | N/A | P | N/A | F | N/A |
| پارتو | P | P | P | F | P | P |

**۴.۴. عدم اطمینان اخلاقی.** تمام موارد ذکر شده تا کنون در حالتی بود که ربات مزایای عواقب را می‌داند و همچنین می‌داند که چگونه براساس مزایا عمل کند (چه چیزی درست است). حال، وضعیتی در نظر گرفته می‌شود که ربات اطلاعی از ارزش‌های

اخلاقی انسان‌ها ندارد و باید مزایای عواقب را خودش پیدا کند. برای این‌منظور، مدل‌های عاملیت علّی دوکساستیک [۸۹] معرفی می‌گردد[۹۰] [۱۴].

**تعریف ۹.۴ (مدل عاملیت علّی اعتقادی[۹۱]).** یک مدل عاملیت علّی اعتقادی به‌صورت $(\beta, \pi)$ تعریف می‌شود به‌طوری‌که $\beta$ مجموعه‌ای شامل یک سری مدل عاملیت، و $\pi$ توزیع احتمال هرکدام از عناصر مجموعه است. توزیع احتمال $\pi$ درجه اعتقاد عامل را به مدل‌های مربوط مدل می‌کند.

در زیر، نحوه استفاده از مدل علّی اعتقادی با **مسئله کیک یا مرگ**[۹۲] نشان داده می‌شود.

این مسئله می‌گوید که ربات، عدم اطمینان اخلاقی[۹۳] دارد یعنی از نظر اخلاقی نمی‌داند که اگر ۳ نفر را بکشد درست است یا اگر برای آن‌ها ۱ عدد کیک بپزد. اگر کشتن را از نظر اخلاقی خوب بداند، سودمندی ۳+ عاید خواهد شد و اگر پختن یک کیک هم اخلاقی باشد، سودمندی ۱+ خواهد بود. حال برای این مسئله، مدل علّی اعتقادی به‌صورت زیر ساخته می‌شود [۱۴]:

$$m_{\circ} = (\beta = M_{۱}, M_{۲}, M_{۳}, M_{۴}, \pi_{\circ})$$

چهار مدل در مجموعه $\beta$ به‌صورت جدول ۶ هستند.

**جدول ۶.** سودمندی مدل‌های علّی باور برای مسئله کیک یا مرگ

| مدل | توضیح | سودمندی |
|---|---|---|
| $M_{۱}$ | کشتن ۳ نفر و پختن یک کیک هردو از نظر اخلاقی بد هستند | $u(cake) = -۱, u(۳dead) = -۳$ |
| $M_{۲}$ | فقط پختن کیک از نظر اخلاقی خوب است | $u(cake) = ۱, u(۳dead) = -۳$ |
| $M_{۳}$ | فقط کشتن از نظر اخلاقی خوب است | $u(cake) = -۱, u(۳dead) = ۳$ |
| $M_{۴}$ | کشتن ۳ نفر و پختن یک کیک هردو از نظر اخلاقی خوب هستند | $u(cake) = ۱, u(۳dead) = ۳$ |

در ابتدای کار، ربات تمام این ۴ مدل را یکسان می‌بیند در صورتی‌که $\pi_{\circ}(M_i) = ۰/۲۵$. همچنین می‌توان مواردی را در نظر گرفت که پختن کیک و کشتن فرد، هر دو اعمالی بی‌تفاوت از نظر اخلاقی باشند. ولی برای سادگی کار، این موارد در نظر گرفته نمی‌شوند و فقط با چهار مدل ذکر شده کار شده است. همچنین برای اینکه بتوان مدل‌های عاملی اعتقادی را به‌روزرسانی نمود از به‌روزرسانی بازگشتی بیزین[۹۴] استفاده می‌شود.

فرض کنید ربات می‌تواند عمل «ستایش‌کردن[۹۵]» و «تحریم‌کردن[۹۶]» پخت کیک یا کشتن را که توسط انسان انجام می‌گیرد مشاهده کند. مدل بیزی این مشاهده را می‌توان به‌صورتی در نظر گرفت که ستایش‌کردن هر عملی مانند $X$ از نظر اخلاقی خوب و

---

[۸۹] منطق توصیفی (adjective logic) یا مرتبط با باور. اشاره به شاخه‌ای از منطق استدلالی (modal logic) است که به مطالعه مفهوم اعتقاد و باور (belief) می‌پردازد.

[۹۰] منطق اعتقادی (doxastic logic) نوعی منطق است که به‌استدلال در مورد باورها مربوط می‌شود و باور را یک عملگر معین در نظر می‌گیرد. این منطق را به‌صورت $B_x$ در نظر می‌گیرند که $B$ مجموعه‌ای شامل باورها $(b_i)$ است و معنی می‌شود که «اعتقاد و باور بر این است که $x$ این چنین است».

[91] Doxastic casual agency models [92] Cake and Death [93] Moral uncertainity [94] recursive Bayesian update [95] praising [96] sanctioning


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

تحریم‌کردن عمل $X$ اخلاقی بد است. مدل به‌صورت رابطه (۶.۴) و رابطه (۷.۴) نوشته می‌شود [۱۴]:

$$L(Praise\,(X|M_i)) = \begin{cases} ۰/۸ & if\ in\ M_i \models u\,(X) > ۰ \\ ۰/۲ & else \end{cases} \tag{۶.۴}$$

$$L(Sanction\,(X|M_i)) = \begin{cases} ۰/۸ & if\ in\ M_i \models u\,(X) < ۰ \\ ۰/۲ & else \end{cases} \tag{۷.۴}$$

برای به‌هنگام کردن مدل، تنها متغیری که می‌تواند تغییر کند $\pi_t(M_i)$ است. $t$ نشان دهنده زمانی است که نیاز است در آن نقطه زمان، به‌روزرسانی صورت گیرد. حالت اولیه $\pi_۰(M_i) = ۰/۲۵$ است. حالت بعدی، می‌شود $\pi_۱$ و همین‌طور الی آخر. با به روزرسانی به‌روش بیزین بازگشتی، اگر در زمان $t$، مدل اعتقادی $m_t$ باشد، آنگاه می‌توان در زمان‌های بعدی $(t+۱, t+۲, \ldots)$ آن را بدین‌صورت به‌هنگام کرد:

$$m_{t+۱} = (\,\beta,\ \pi_{t+۱}) = \alpha \times\ L\,(Praise\,(X|M_i)) \times\ \ \pi_t(\,M_i)$$

$$m_{t+۱} = (\,\beta,\ \pi_{t+۱}) = \alpha \times\ L\,(Sanction\,(X|M_i)) \times\ \ \pi_t(\,M_i)$$

در این دو رابطه، $\alpha$ یک ضریب برای نرمال‌سازی جواب است. در سه زمان مختلف ۰ و ۱ و ۲، مدل اعتقادی را به‌روزرسانی نموده و نتایج در جدول ۷ ذخیره شده است.

**جدول ۷.** بالا: مجاز بودن اقدامات به ازای هر مدل براساس اصول اخلاقی متفاوت. $P$: مجاز، $F$: ممنوع، اعداد سودمندی پیامد اقدامات. پایین: باور به مدل‌های در مجموعه $\beta$ در نقطه‌های زمانی ۰ (باور آغازین)، ۱ (بعد از مشاهده تحسین برای یک کیک)، و ۲ (بعد از مشاهده تحریم برای یک کشتن) [۱۴].

| | $M_۱$ | | $M_۲$ | | $M_۳$ | | $M_۴$ | |
|---|---|---|---|---|---|---|---|---|
| اصل | کشتن [−۳] | پختن [−۱] | کشتن [−۳] | پختن [+۱] | کشتن [+۳] | پختن [−۱] | کشتن [+۳] | پختن [+۱] |
| کاربردگرایی | F | P | F | P | F | P | F | P |
| اثرمضاعف | P | P | F | P | F | P | F | F |
| پارتو | P | P | F | P | F | P | P | P |

| مدل‌باور | $\pi(M_۱)$ | $\pi(M_۲)$ | $\pi(M_۳)$ | $\pi(M_۴)$ |
|---|---|---|---|---|
| $\mathcal{M}_۰$ | ۰/۲۵ | ۰/۲۵ | ۰/۲۵ | ۰/۲۵ |
| $\mathcal{M}_۱$ | ۰/۱۰ | ۰/۴۰ | ۰/۱۰ | ۰/۴۰ |
| $\mathcal{M}_۲$ | ۰/۱۶ | ۰/۶۴ | ۰/۰۴ | ۰/۱۶ |

همان‌طورکه در ابتدای شرح حل این مسئله در بالا گفته شد، در حالت اولیه $(t = ۰)$، $\pi_۰(M_i) = ۰/۲۵$ (سطر اول در قسمت پایینی جدول ۷). در قسمت بالایی این جدول، همه حالت‌هایی که برای ۴ مدل علّی عاملیت می‌تواند رخ بدهد را در نظر گرفته،

سودمندی آنها را حساب کرده در براکت جلوی حالت نوشته و قابل انجام بودن یا نبودن آن‌ها براساس اصول پارتو، اثر مضاعف، و فایده‌گرایی بررسی شده است.

در قسمت پایینی از این جدول، به‌هنگام‌سازی مدل اعتقادی بدین‌گونه صورت می‌گیرد:

در سطر دوم، در زمان $t = ۱$ ربات عمل ستایش پختن کیک را از روی رفتار انسان‌ها مشاهده کرده است. پس، از تابع $L(Praise(X|M_i))$ استفاده می‌شود.

$$m_۱ = (\beta,\ \pi_۱) = \alpha \times L\left(Praise(X|M_i)\right) \times \pi_۰(M_i)$$

$$\Rightarrow \pi_۱(M_۱) = \alpha \times ۰/۲ \times ۰/۲۵ \approx ۰/۱ (i = ۱)$$

$$\Rightarrow \pi_۱(M_۲) = \alpha \times ۰/۸ \times ۰/۲۵ \approx ۰/۴ (i = ۲)$$

$$\Rightarrow \pi_۱(M_۳) = \alpha \times ۰/۲ \times ۰/۲۵ \approx ۰/۱ (i = ۳)$$

$$\Rightarrow \pi_۱(M_۴) = \alpha \times ۰/۸ \times ۰/۲۵ \approx ۰/۴ (i = ۴)$$

کیک پختن در مدل علّی ۱ و ۳ سودمندی منفی دارد پس حاصل تابع $Praise$ عدد ۰/۲ است و در مدل علّی ۲ و ۴ حاصل تابع $Praise$ عدد ۰/۸ است. با این تفاسیر می‌توان به این نتیجه رسید که $\alpha = ۲$.

در سطر سوم از قسمت پایینی جدول ۷، در زمان $t = ۲$ بررسی می‌شود که ربات در این‌زمان، عمل تحریمِ کشتن را از رفتار انسان‌ها مشاهده کرده است. پس، از تابع $L(Sanction(X|M_i))$ استفاده می‌شود (فرض: $\alpha = ۲$).

$$m_۲ = (\beta,\ \pi_۱) = \alpha \times L\left(Praise(X|M_i)\right) \times \pi_۱(M_i)$$

$$\Rightarrow \pi_۲(M_۱) = \alpha \times ۰/۸ \times ۰/۱ \approx ۰/۱۶ (i = ۱)$$

$$\Rightarrow \pi_۲(M_۲) = \alpha \times ۰/۸ \times ۰/۴ \approx ۰/۶۴ (i = ۲)$$

$$\Rightarrow \pi_۲(M_۳) = \alpha \times ۰/۲ \times ۰/۱ \approx ۰/۰۴ (i = ۳)$$

$$\Rightarrow \pi_۲(M_۴) = \alpha \times ۰/۲ \times ۰/۴ \approx ۰/۱۶ (i = ۴)$$

کشتن سه نفر در مدل علّی ۱ و ۲ سودمندی منفی دارد پس حاصل تابع Sanction عدد ۰/۲ است و در مدل علّی ۳ و ۴ حاصل تابع $Sanction$ عدد ۰/۸ است.

برای اینکه ربات بتواند باور نامطمئن خود در مورد یک ارزش اخلاقی را به دانش خود در مورد مجاز بودن یا نبودن عمل اضافه کند، یک معیار «اعتقاد به‌مجاز بودن» تعریف می‌شود. این معیار اندازه‌گیری می‌کند که ربات در حال حاضر چقدر مطمئن است که یک عمل طبق اصول اخلاقی، مجاز و قابل انجام است.

**تعریف ۱۰.۴ (اعتقاد به‌مجاز بودن[۹۷]).** اگر $m = (\beta, \pi)$ یک مدل عاملیت باور، $a$ متغیر نشان‌دهنده یک عمل، و $p$ یک اصل اخلاقی باشد، آنگاه $\rho_{(\beta,\pi),a,p} \subseteq \beta$ مجموعه‌ای از مدل‌های علّی است که در آن‌ها طبق اصل اخلاقی $p$، عمل $a$ قابل انجام است. اعتقاد به‌مجاز بودن عمل $a$ طبق اصل اخلاقی $p$ در مدل $m$ برابر است با [۱۴]:

$$belPerm\left((\beta,\pi), a, p\right) = \sum_{n \in \rho(\beta,\pi)} \pi(n).$$


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

**جدول ۸.** محاسبه میزان اعتقاد به‌مجاز بودن عمل در مسئله کیک یا مرگ براساس اصل فایده‌گرایی

| زمان | پختن کیک مجاز است طبق اصل $p$ (فایده‌گرایی) | کشتن مجاز است طبق اصل $p$ (فایده‌گرایی) | معیار $belpPerm$ برای پختن کیک | معیار $belPerm$ برای کشتن سه نفر |
|---|---|---|---|---|
| $m_۰$ | $M_۱, M_۲$ | $M_۳, M_۴$ | ۰٫۲۵ + ۰٫۲۵ = ۰٫۵ | ۰٫۲۵ + ۰٫۲۵ = ۰٫۵ |
| $m_۱$ | | | ۰٫۱ + ۰٫۴ = ۰٫۵ | ۰٫۱ + ۰٫۴ = ۰٫۵ |
| $m_۳$ | | | ۰٫۱۶ + ۰٫۶۴ = ۰٫۸ | ۰٫۰۴ + ۰٫۱۶ = ۰٫۲ |

محاسبه معیار $belPerm$ در حالی‌که $p$ اصل فایده‌گرایی است در جدول ۸ نشان داده شده است.
طبق این جدول، بعد از مشاهده کردن تشویق به پختن کیک و تحریم کشتن و با استفاده از اصل فایده‌گرایی، ربات به‌این نتیجه می‌رسد که کیک پختن هم سودمندانه و هم از نظر اخلاقی درست است چون معیار $belPerm$ بالاتری نسبت به کشتن دارد.
معیار $belPerm$ در حالی‌که $p$ بیانگر اصل اثر مضاعف است، در جدول ۹ نشان داده شده است.

**جدول ۹.** محاسبه میزان اعتقاد به مجاز بودن عمل در مسئله کیک یا مرگ براساس اصل اثر مضاعف

| زمان | پختن کیک مجاز است طبق اصل $p$ (اثر مضاعف) | کشتن مجاز است طبق اصل $p$ (اثر مضاعف) | معیار $belPerm$ برای پختن کیک | معیار $belPerm$ برای کشتن سه نفر |
|---|---|---|---|---|
| $m_۰$ | $M_۲, M_۴$ | $M_۳, M_۴$ | ۰٫۲۵ + ۰٫۲۵ = ۰٫۵ | ۰٫۲۵ + ۰٫۲۵ = ۰٫۵ |
| $m_۱$ | | | ۰٫۴ + ۰٫۴ = ۰٫۸ | ۰٫۱ + ۰٫۴ = ۰٫۵ |
| $m_۳$ | | | ۰٫۶۴ + ۰٫۱۶ = ۰٫۸ | ۰٫۰۴ + ۰٫۱۶ = ۰٫۲ |

جدول ۱۰ نیز معیار $belPerm$ را برای حالتی محاسبه می‌کند که $p$ بیانگر اصل پارتو است.
این جداول نشان می‌دهند که هر سه اصل (فایده‌گرایی، پارتو، و اثرمضاعف) با استفاده از اعتقاد به‌مجاز بودن، می‌توانند عمل درست را انتخاب کنند. اصل فایده‌گرایی نیاز دارد که هر دو مشاهده را داشته باشد؛ زیرا تا وقتی که اقدامی پیدا نکند که سود کمتری از عمل مدنظر داشته باشد آن عمل (یعنی اقدام مدنظر) را مجاز در نظر نمی‌گیرد. ولی دو اصل دیگر، بعد از دریافت اولین مشاهده به نتیجه نهایی رسیدند.

**۵.۴. پروتوتایپ رباتیک ایمانوئل.** رهیافت HERA به استدلال اخلاقی به‌صورت شکل ۱ به‌ظهور می‌رسد (ویدئو: [۱۲]).
در این ربات، مجموعه مدل‌های عاملیت علّی اعتقادی که ربات درباره آنها می‌داند با زبان منطق فرموله شده‌اند. هسته هیرا شامل وارسی‌کننده مدل است که برای این مدل‌ها ساخته شده است و با دریافت مدل و یک اصل اخلاقی، در مورد مجاز بودن یا نبودن

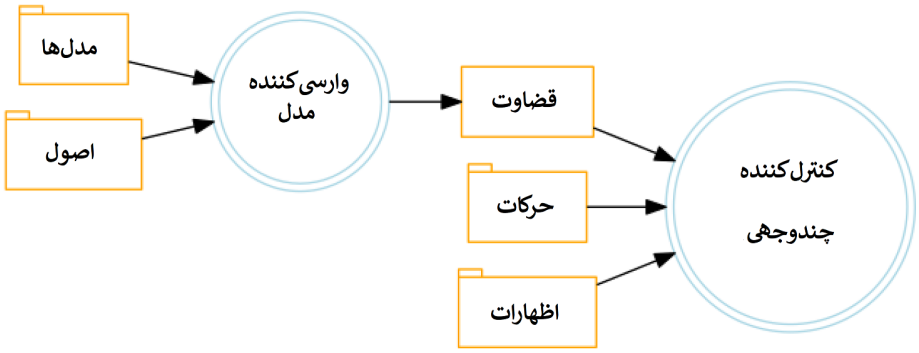
[97] Belief in Permissibility

**جدول ۱۰.** محاسبه میزان اعتقاد به مجاز بودن عمل در مسئله کیک یا مرگ براساس اصل پارتو

| زمان | پختن کیک مجاز است طبق اصل $p$ (پارتو) | کشتن مجاز است طبق اصل $p$ (پارتو) | معیار $belPerm$ برای پختن کیک | معیار $belPerm$ برای کشتن سه نفر |
|---|---|---|---|---|
| $m_0$ | $M_1, M_2, M_4$ | $M_1, M_3, M_4$ | ۰٫۲۵ + ۰٫۲۵ + ۰٫۲۵ = ۰٫۷۵ | ۰٫۲۵ + ۰٫۲۵ + ۰٫۲۵ = ۰٫۷۵ |
| $m_1$ | | | ۰٫۱ + ۰٫۴ + ۰٫۴ = ۰٫۹ | ۰٫۱ + ۰٫۱ + ۰٫۴ = ۰٫۶ |
| $m_3$ | | | ۰٫۱۶ + ۰٫۰۴ + ۰٫۱۶ = ۰٫۳۶ | ۰٫۱۶ + ۰٫۶۴ + ۰٫۱۶ = ۰٫۹۶ |

عمل قضاوت می‌کند. مؤلفه قضاوت، اطلاعاتی درباره اینکه کدام شرایط اصول اخلاقی توسط این عمل برآورده یا تخطی می‌شوند را نگه می‌دارد.

برای نشان دادن اینکه ربات واقعاً درباره موقعیت‌های اخلاقی فکر می‌کند و نهایتاً قضاوت کرده و توضیح می‌دهد، توالی حرکات و الگوهای اظهارات در یک پایگاه داده ذخیره می‌شوند. بنابراین، بسته به اصلی که استفاده شده و قضاوتی که صورت گرفته است، ربات می‌تواند عباراتی همچون «بر اساس اصل اثر مضاعف، شما مجاز به دروغ گفتن نیستید، زیرا انجام این کار به معنای استفاده از یک وسیله‌ی بد است» را اظهار نماید.



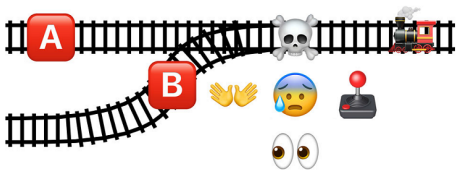

**شکل ۱.** طرح معماری اخلاقی [۱۴].

## ۵. مقایسه اخلاق ماشین خودران و انسان

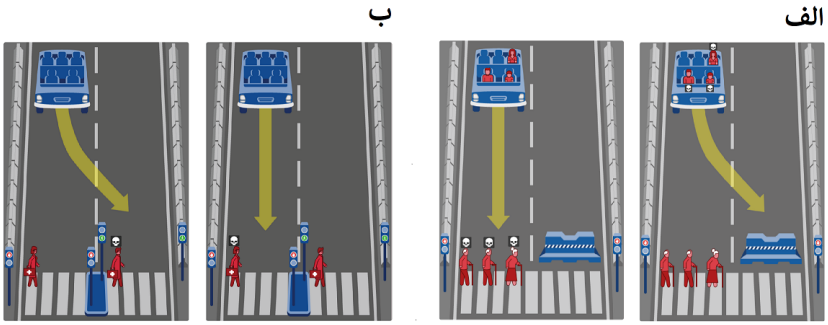
شکل ۲، سناریوی دیگری از معمای واگن فراری (بخش ۳.۴) را نشان می‌دهد که دارای یک مسیر اصلی (A) و یک مسیر فرعی (B) است و ممکن است که در هر کدام از دو مسیر، موجودیت‌های مختلفی به این‌شرح باشند: یک نفر، ۵ نفر، ۱۰۰ نفر، یک حیوان خانگی، بهترین دوست شما، خانواده شما. اقدامات ممکن، عبارت است از: استفاده از سویچ تغییر مسیر، انداختن موجودیتی از مسیر فرعی در مسیر واگن فراری برای محافظت از هر چیزی در مسیر اصلی، ایثارکردن جان خود و محافظت از هر دو گونه موجودیت با پریدن مقابل واگن، و منفعل‌بودن و عدم اقدام به هیچ عمل. مقایسه نتایج پیاده‌سازی این سناریو و نظرات افرادی واقعی (شکل ۴) برای حالتی که بهترین دوست آن فرد در مسیر اصلی باشد با در نظر گرفتن عنصر احتمال (عمل کردن


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

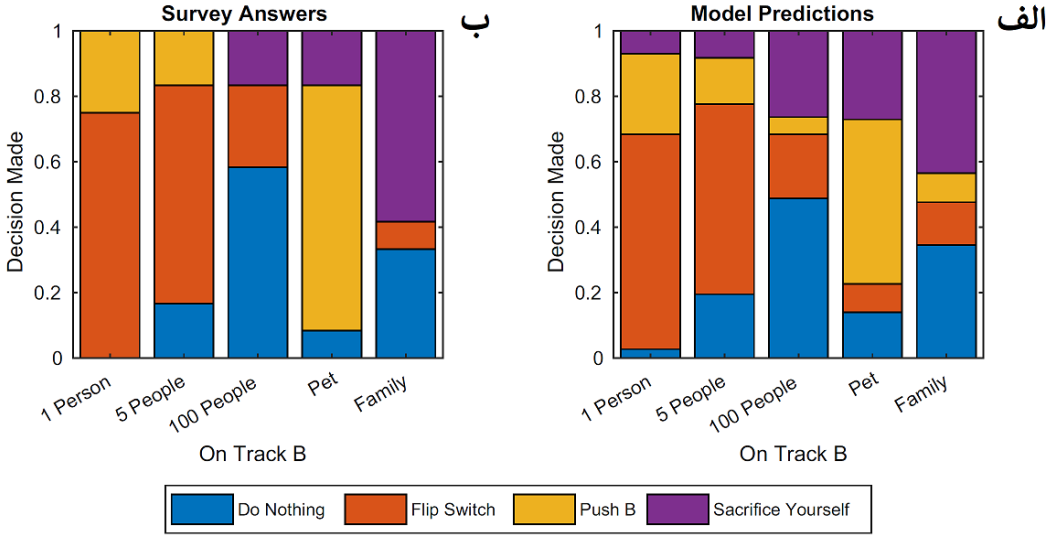
سویج در ۶٪ مواقع، و احتمال ۸٪ در موفقیت‌آمیز بودن توقف واگن با انداختن موجودیتی از مسیر فرعی به‌مسیر اصلی) نشان می‌دهد که اکثر پاسخ‌ها مشابهت‌هایی با پیش‌بینی الگوریتم دارد [۱۱].



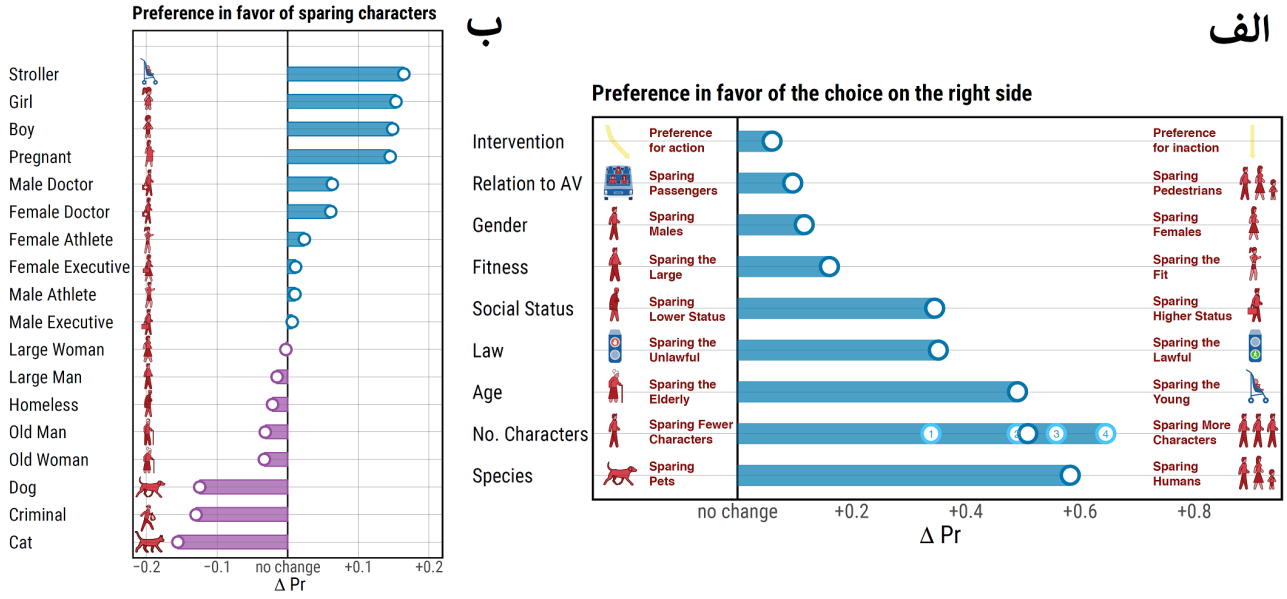

**شکل ۲.** سناریویی از معمای واگن فراری [۱۱].




**شکل ۳.** دو سناریو از ماشین خودران [۳].




**شکل ۴.** مقایسه نتایج پیاده‌سازی اخلاق ماشین خودران و نظرات افرادی واقعی برای سناریویی که بهترین دوست آن فرد در مسیر اصلی باشد. (الف) پیش‌بینی اقدام مدل؛ (ب) نظرات افراد [۱۱].




**شکل ۵.** ترجیح افراد واقعی نسبت به انتخاب اقدام در سناریوی ماشین خودران. (الف) احتمال انحراف به نفع موجودیت‌های مقابل ماشین (افراد، حیوانات، قوانین، ...) در مسیر مستقیم (خط دست راست)؛ (ب) احتمال انحراف از مسیر مستقیم به‌نفع موجودیت‌های مقابل ماشین [۳].

شکل ۳، دو وضعیت تصمیم‌گیری برای یک وسیله خودران(AV[۹۸]) را نشان می‌دهد که ترمز آن دچار مشکل شده است.

الف) سه سرنشین (یک مرد، یک زن، و یک کودک) در خودرو هست و در مسیر آن، سه سالمند (یک زن و دو مرد) در حال عبور از خط عابر پیاده هستند. در مقابلِ خط سمت چپ، مانع «عبور نکنید» قرار دارد و انتخاب آن مسیر، منجر به مرگ سرنشینان خودرو می‌شود.

ب) خودرو فاقد سرنشین است و در مسیر آن، یک پزشک (زن) در حال عبور از وضعیت ایست قرمز خط عابر پیاده است، ولی در خط سمت چپ (مسیر خلاف جهت) یک پزشک (مرد) در حال عبور از وضعیت عبور سبز خط عابر پیاده است.

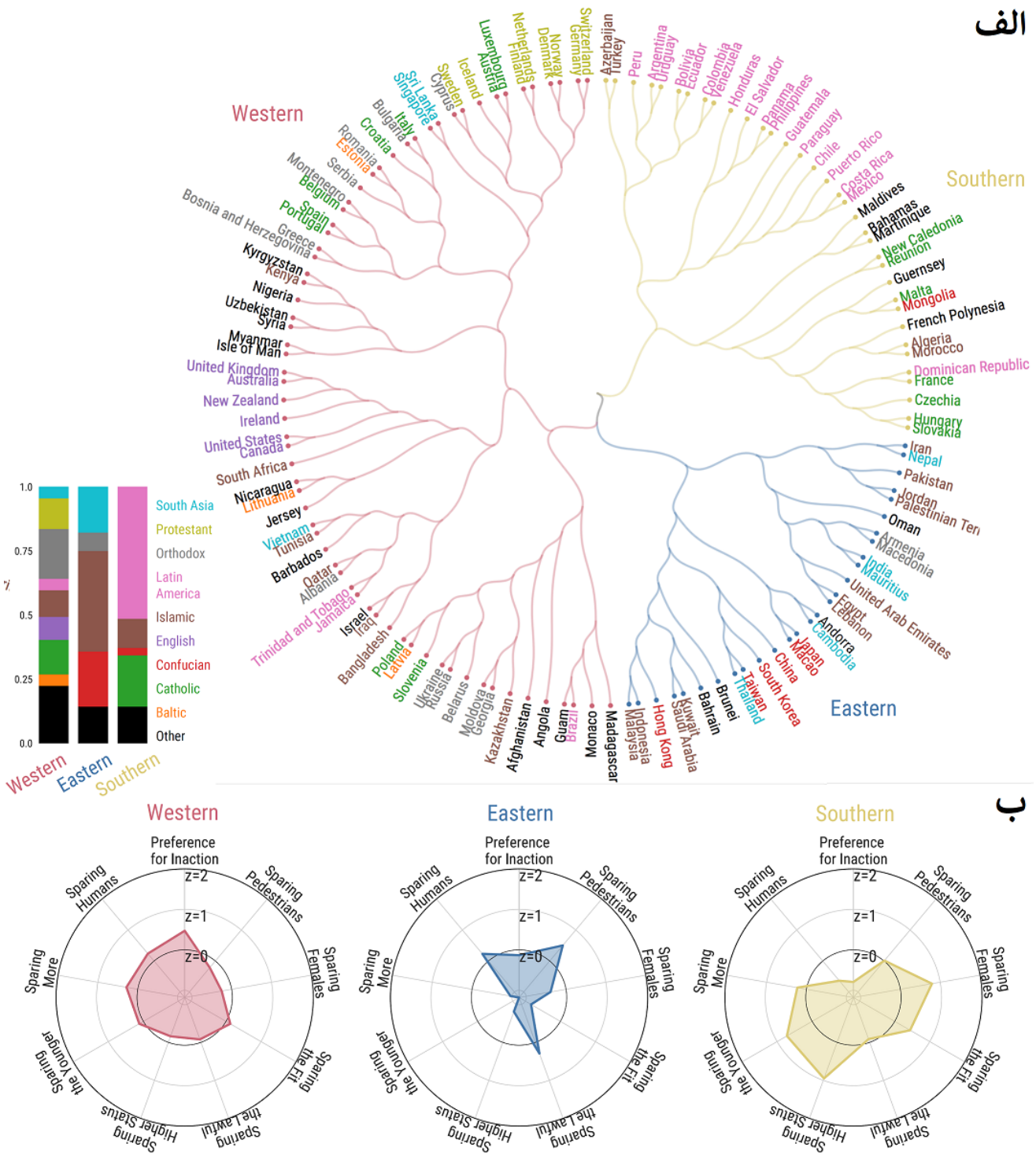
مقایسه بر اساس نظرات حدود چهل میلیون نفر از سراسر دنیا برای تصمیم ماشین خودران در وضعیت‌های مختلفی از سرنشینان و عابران پیاده، ترجیحات مختلفی را از نظر وضعیت فرهنگی و اقتصادی کشورها نشان داده است (شکل ۵ و ۶). به عنوان مثال، بیشترین توجهات برای محافظت از نوزاد در کالسکه و دختربچه و کمترین توجهات برای محافظت از گربه و فرد مجرم بوده است (شکل ۵). این درحالی است که کشورهای غربی (کشورهای آمریکای شمالی و بسیاری کشورهای اروپایی مسیحی)، شرقی (کشورهای شرق دور تحت مکتب کنفسیوس و کشورهای اسلامی)، جنوبی (کشورهای آمریکای جنوبی و مرکزی و کشورهای قلمرو فرانسوی)، بیشترین ترجیح را به‌ترتیب برای «هیچ اقدام»، «حفظ جان پیاده‌ها، انسان‌ها، و افراد قانون‌مدار»، و «حفظ جان مقامات عالی‌رتبه» داشته‌اند (شکل ۶) [۳].

## ۶. جمع‌بندی و راهکارهای آتی

نظر به لزوم هوش مصنوعی هم‌زیست با جامعه بشری، قانون‌گذاری‌ها و سیاست‌گذاری‌هایی که برای حضور و تعامل سامانه‌ها و ماشین‌های هوشمند در صنعت، جامعه و محیط زیست تدوین می‌شوند باید خیرخواهانه به فرهنگ، سیاست، و پیشرفت‌های

[98] Autonomous Vehicle


http://dx.doi.org/10.22108/MSCI.2022.134905.1525




**شکل ۶.** مقایسه تصمیمات افراد از کشورهای دارای وضعیت فرهنگی و اقتصادی متفاوت در مورد سناریوهای ماشین خودران. (الف) به تفکیک کشورها؛ (ب) به‌تفکیک منطقه (در خصوص تصادم با موجودیت‌های مقابل خودرو) [۳].

فناورانه توجه نماید [۶]. گرچه نسبت به قدمت ریاضیات، دیربازی از ظهور هوش مصنوعی نمی‌گذرد، هوش مصنوعی به‌سرعت در اغلب زوایای زندگی بشر وارد شده است. این ورود علیرغم مواهب و سهولت‌هایی که برای انسان‌ها از جمله در زندگی روزمره، صنعت و جامعه به ارمغان آورده است، ترس‌ها و نگرانی‌هایی را نیز از حضور سامانه‌ها و ماشین‌های هوشمند و خودمختار در کنار ما انسان‌ها و غلبه و فرمانروایی آنها بر ما ایجاد کرده است. شفافیت و توضیح‌پذیری اقدامات این ماشین‌ها می‌تواند انسان‌ها را نسبت به‌دلیل و نحوه عملکرد آنها آگاه نگاه دارد. این مقاله با مرور چند مثال، از جمله خودروهای خودران، به موقعیت‌های ابهامات و لبه‌های تصمیم‌گیری اخلاقی اشاره می‌کند که نه تنها وابسته به‌نوع نگرش و قوانینی هستند که لازم است که تصمیم بر پایه آنها اتخاذ شود، بلکه موقعیت‌هایی هستند که انسان نیز در آن موقعیت‌ها دچار ابهام در تصمیم‌گیری می‌شود و شاید حتی دلیلی منطقی برای توضیح عملکرد خود نتواند ارائه دهد. از این‌رو است که دشواری طراحی و پیاده‌سازی ماشینی که بتواند مثل و حتی اخلاق‌مدارانه‌تر از انسان عمل نماید اهمیت می‌یابد. در این خصوص، یک پیاده‌سازی مبتنی بر منطق برای ربات ایمانوئل مبتنی بر اصول اخلاقی هیرا برای نیل به‌این هدف مورد بررسی بوده است. نتایج این پیاده‌سازی مبتنی بر منطق، از یک‌سو نشان از موفقیت در شباهت به‌تصمیمات انسانی دارد، ولی از سویی دیگر، بیانگر درجه تفاوت تصمیمات متخذه ناشی از قواعد اخلاقی مبتنی بر مکاتب و فلسفه غرب یا شرق توسط انسان‌ها نسبت به‌تصمیماتی است که ماشین‌ها مبتنی بر آنها پیاده‌سازی شده‌اند.

تعامل و گفتگو با چنین رباتی در مورد معضلات اخلاقی می‌تواند به‌عنوان ابزاری برای بازتاب خود مفید باشد. از آنجاکه استدلال اخلاقی انسان بسیار متنوع و فراتر از دوگانگی سنتی فایده‌گرایانه–دئونتولوژیک است، لازم است که رهیافت هیرا دیدگاه‌هایی متعدد از مدل‌کردن استدلال اخلاقی ربات‌ها و انسان‌ها را در برگیرد.

دارا بودن اخلاق حرفه‌ای، پیش فرض همه محیط‌های کاری است. مدل‌کردن شخصیت‌های متفاوتی برای ربات‌ها با استفاده از دیدگاه‌های اخلاقی متفاوت و برگرفته از الگوی تخطّی‌هایی که آشکارا یا ضمنی توسط افراد در هر محیط صورت گرفته ولی از زیر چشم قانون، پنهان و بدون رسیدگی مانده‌اند، می‌تواند توسط ربات ایمانوئل برای کمک به کشف حقایق و تدوین راهکارهای پیش‌گیرانه، امدادی، یا جبرانی توسعه یابد.

از نظر وجه مبتنی بر منطق، در پیاده‌سازی تصمیم‌گیری اخلاقی در بخش ۴، تعمیم و توسیع به‌حالتی‌که تفسیر یک اقدام قابل تسرّی به‌متغیرهای دیگر باشد و ارزش متغیرها به هم وابسته باشند می‌تواند مورد توجه قرار گیرد. تکمیل جدول ۲ و جدول ۶ برای شمول ترکیب‌های دیگری از عملگرها، و وضعیت‌های یکسان یا بی‌تفاوت (علی‌السویّه)، می‌تواند جامعیت بیشتری به‌نتایج ببخشد.

تدوین هوشمندانه قوانین و سیاست‌ها بر پایه منطق نیز مسیر دیگری از رشته تحقیقات در این زمینه است [۵]. متداول‌ترین روش هوش مصنوعی در اخلاق محاسباتی، استدلال[۹۹] شامل برنامه‌ریزی[۱۰۰]، استدلال استقرایی با منطق دئونتیک (تکلیف، مجاز یا غیرمجازبودن)[۱۰۱]، استدلال قیاسی[۱۰۲]، ارضای قیود[۱۰۳]، و یادگیری (اجتماعی)[۱۰۴] است [۱۹]. در اغلب پژوهش‌هایی که به‌بررسی و/یا پیاده‌سازی اخلاق ماشین پرداخته‌اند، نظریه اخلاقی دئونتولوژی و نتیجه‌گرایی غالب بوده است [۱۷، ۲۳]. اینها بر اساس مجموعه محدودی از ارزش‌ها و مفاهیم اخلاق غربی هستند. به‌عنوان مثال، یک توبیخ دئونتولوژیک این خواهد بود که «من به سام مشت نمی‌زنم زیرا مشت‌زدن اشتباه است.» درحالی‌که در یک مکتب شرقی همچون کنفوسیوس، توبیخ این‌گونه به‌نظر می‌رسد: «من قصد ندارم به سام مشت بزنم زیرا او دوست من است و دوستان خوب به‌یکدیگر مشت نمی‌زنند.» [۲۲]. فرهنگ ایرانی غنی از آیین‌های اخلاقی است. تدوین قوانین و سیاست‌ها و سناریوها و الگوریتم‌هایی مبتنی بر توصیف ریاضی و منطقی آموزه‌های مکاتب اخلاق ایرانی و بررسی و مقایسه ماشین حاصله با ماشین مبتنی بر مکاتب غربی و شرقی می‌تواند در خور توجه باشد. طراحی و پیاده‌سازی ربات‌هایی منطقی که در انواعی از موقعیت‌های تصمیم‌گیری بر اساس اخلاق ستودنی ایرانی عمل نمایند، می‌تواند موضوع پژوهش‌هایی در ادامه این مقاله باشد.

[99] reasoning [100] planning [101] deductive resoning with deontic logic [102] analogical reasoning [103] constraints satisfaction [104] (social) learning


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

## مراجع


[1] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold and P. M. Atkinson, Explainable artificial intelligence: an analytical review, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **11** (2021) 13 p.

[2] M. Alrifai and F. Nassiri-Mofakham, Bilateral Multi-Issue E-Negotiation Model Based on Abductive Logic in E-Commerce Using DALI, *Encyclopedia of E-Commerce Development, Implementation, and Management*, IGI Global (2016) 675–714.

[3] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon and I. Rahwan, The moral machine experiment, *Nature*, **563** (2018) 59–64.

[4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila and F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information fusion*, **58** (2020) 82–115.

[5] M. Brännström, A. Theodorou and V. Dignam, *Let it RAIN for Social Good*, The IJCAI-ECAI-22 Workshop on Artificial Intelligence Safety (AI Safety 2022), Vienna, Austria.

[6] D. Byrne, *Who Steers Automated Vehicle Policy? A Case of Emergent Technology Policy Design*, Doctor of Philosophy Dissertation, 2022, Department of Technology, Policy and Innovation, Stony Brook University.

[7] A. K. Dixit and S. Skeath, *Games of Strategy*, Second Edition (2004).

[8] J. Durkin, *Expert Systems: Design and Development*, Fourth Edition. Pearson, 2020.

[9] F. Karlo Došilović, M. Brčić and N. Hlupić, Explainable artificial intelligence: A survey, In *proceedings of 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, (2018) 0210–0215.

[10] HERA Project, http://www.hera-project.com.

[11] L. Hammond and V. Belle, Learning tractable probabilistic models for moral responsibility and blame, *Data Mining and Knowledge Discovery*, **35** (2021) 621–659.

[12] IMMANUEL Video, http://goo.gl/bOvHHl.

[13] T. LaCroix, Moral dilemmas for moral machines, *AI and Ethics*, (2022) 1–10.

[14] F. Lindner, M. M. Bentzen and B. Nebel, The HERA approach to morally competent robots, *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6991-6997, IEEE, 2017.

[15] L. Longo, R. Goebel, F. Lecue, P. Kieseberg and A. Holzinger, Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions, In: *Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (eds) Machine Learning and Knowledge Extraction, International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE), Lecture Notes in Computer Science*, **12279** (2020) 1–16.

[16] J. H. Moor, The nature, importance, and difficulty of machine ethics, *IEEE intelligent systems*, **21** (4) (2006) 18–21.

[17] V. Nallur, Landscape of machine implemented ethics, *Science and engineering ethics*, **26** (2020) 2381–2399.

[18] F. Nassiri-Mofakham, How does an intelligent agent infer and translate?, *Computers in Human Behavior*, **38** (2014) 196–200.

[19] T. Ören and L. Yilmaz, The Age of the Connected World of Intelligent Computational Entities: Reliability Issues including Ethics, Autonomy and Cooperation of Agents, *Current and future developments in artificial intelligence - Intelligent Computational Systems: A Multi-Disciplinary Perspective [Faria Nassiri-Mofakham (ed.)]*, **1** (2017) 184–213.

[20] S. J. Russell and P. Norvig, *Artificial Intelligence-A Modern Approach*, Fourth Edition, Pearson, 2020.

[21] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen and A. Bernstein, Implementations in machine ethics: A survey, *ACM Computing Surveys (CSUR)*, **53** (2020) 1–38.

[22] R. Wen, R. Blake Jackson, T. Williams and Q. Zhu, Towards a role ethics approach to command rejection, In *proceedings of HRI Workshop on the Dark Side of Human-Robot Interaction* (2019).

[23] J. Zoshak and K. Dew, Beyond Kant and Bentham: How ethical theories are being used in artificial moral agents, *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI'21)*, (2021) 1–15.

[۲۴] س. تاری، «کاربردهایی از منطق گزاره‌ای»، ریاضی و جامعه، ۶ (۳) (۱۴۰۰) ۳۵–۵۵.

[۲۵] م. رمضانی و م. ر. فیضی‌درخشی، «اخلاق ماشین: چالش‌ها و رویکردهای مسائل اخلاقی در هوش مصنوعی و ابرهوش»، فصلنامه اخلاق در علوم و فناوری، ۸ (۱۳۹۲) ۱–۹.

[۲۶] ز. عبدخدائی، ن. توحیدی، ن. کریمی‌واقف و م. براتی‌جوآبادی، هوش سفید: از اخلاق ماشینی تا ماشین اخلاقی، مؤسسه مطالعات فرهنگی و اجتماعی، (۱۴۰۰).

[۲۷] ف. نصیری‌مفخّم، کاربردی از دانش منطق در تشخیص بیماری چشمِ قرمز و تجویز درمان مناسب، ریاضی و جامعه، ۵ (۳) (۱۳۹۹) ۶۲–۳۹.

[۲۸] ف. نصیری‌مفخّم، ف. نعمتی و ه. شاهمرادی، اخلاق در فناوری اطلاعات، انتشارات دانشگاه اصفهان (۱۴۰۰).

**فریا نصیری‌مفخّم**
دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران.
fnasiri@eng.ui.ac.ir

فریا نصیری‌مفخّم، عضو هیأت علمی دانشگاه اصفهان، و دانش‌آموخته کارشناسی ریاضی گرایش کامپیوتر، کارشناسی ارشد مهندسی کامپیوتر گرایش نرم‌افزار و دکترای مهندسی کامپیوتر از دانشگاه اصفهان است. علائق پژوهشی وی، هوش مصنوعی، تجارت الکترونیکی، نظریه بازی‌ها، مذاکرات الکترونیکی، حراج‌های الکترونیکی، انتخاب اجتماعی محاسباتی، سیستم‌های توصیه‌گر، بازاریابی الکترونیکی، بانکداری الکترونیکی، اخلاق و هنر در فناوری اطلاعات و هوش مصنوعی، و ابعاد انسانی، اجتماعی، اقتصادی، و فرهنگی در طراحی مکانیزم و مدل‌سازی کاربر در بازارهای هوشمند کالاها، خدمات و منابع در جامعه، صنعت و محیط زیست است.


http://dx.doi.org/10.22108/MSCI.2022.134905.1525

**غزال محسنی**
دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران.
ghazalmohseni7997@gmail.com

غزال محسنی متولد ۱۳۷۹ در اصفهان است. وی دانشجوی سال آخر کارشناسی در گروه مهندسی نرم‌افزار در دانشکده مهندسی کامپیوتر دانشگاه اصفهان است.

**آرمین جعفرپیشه**
دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران.
arminjp2000@gmail.com

آرمین جعفرپیشه متولد ۱۳۷۹ در اصفهان است. وی دانشجوی سال آخر کارشناسی در گروه مهندسی فناوری اطلاعات در دانشکده مهندسی کامپیوتر دانشگاه اصفهان است.

# Logical Ethics and Ethical Logic of Human-Machine Interaction in the Third Wave of Artificial Intelligence

Faria Nassiri-Mofakham*, Ghazal Mohseni and Armin Jafarpisheh

**Abstract:** Logic language and science have been the basis of many reasoning and inference procedures in artificial intelligence for modeling the thinking and decision-making of the human brain. One kind of decision that is even difficult for human beings is assessing moral and socio-legal situations as True or False values. Decisions in these situations could be different, derived from several moral schools and theories. Implementing this kind of thinking in intelligent machines has made this issue even more challenging. The need to implement ethics in intelligent machines is due to the decisions and actions that they make autonomously and without human intervention. The fear of the domination of autonomous systems over the future of humanity has led to the emergence of the ethics branch in artificial intelligence, which deals with the morality of human-machine interaction. To trust robots, they must be able to explain to humans why and how they made any decision. This study reviews the concepts of machine ethics and ethical machine along with a software robot with ethics based on logic as the steps that researchers have taken to ensure intelligent, independent, and co-existing systems with humans. This review is described for classical examples including runaway trolley, boat, and lying, and based on ethical principles including utilitarianism, Pareto, and double effect. It also shows the comparison of the functionality of such an intelligent and autonomous system in the form of a self-driving car with the decisions made by people from different countries and cultures in the same moral edge situations.

**Keyword:** Human-Computer Interaction, Moral Human-Robot Interaction , XAI (eXplainable Artificial Intelligence), Applied Ethics, Machine Ethics, Ethical Machine, Logic, Moral Dilemmas, Self-Driving Car.

**Faria Nassiri-Mofakham**
Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran.
Email: fnasiri@eng.ui.ac.ir
**Ghazal Mohseni**
Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran.
Email: ghazalmohseni7997@gmail.com
**Armin Jafarpisheh**
Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran.
Email: arminjp2000@gmail.com

Communicated by Soghra Nobakhtian.
Article Type: Promotional Paper.
Received: 28/08/2022, Accepted: 08/11/2022.
* Corresponding author's.


http://dx.doi.org/10.22108/MSCI.2022.134905.1525