رفتار واقعی هوش‌مصنوعی برای دانشمندان داده

انتشارات محمد رحیمی

۴ ژوئن ۲۰۲۳

فهرست مطالب

تشکر نامه

مایلم از مشارکت کنندگان زیر برای کمک‌های عالی و متنوعشان در این کتاب تشکر کنیم: پیتر هرشوک (اخلاق بودایی)، جان هکر رایت (اخلاق فضیلت)، ساموئل جی لوین و دانیل سینکلر (اخلاق یهودی)، کالین مارشال (اخلاق دئونتولوژیک)، جوی میلر و آندریا سالیوان کلارک (اخلاق بومی) و جان مورانگی (اخلاق آفریقایی). هدف ما ترسیم تصویری اخلاقی است که تا حد امکان منتوع و جذاب باشد. بدون کمک این عزیزان خردمند، توانستیم این کتاب را ایجاد کنیم!

فهرست همکاران و مشارکت کنندگان

همکاران:

پیتر سینگر
مرکز دانشگاهی برای ارزش‌های انسانی دانشگاه پرینستون

ییپ فای تسه
مرکز دانشگاهی برای ارزش‌های انسانی دانشگاه پرینستون

ساموئل جی لوین
مرکز حقوقی تورو

کولین مارشال
گروه فلسفه دانشگاه واشنگتن

جویی میلر
گروه فلسفه دانشگاه وست چستر

جان مورونگی
گروه فلسفه دانشگاه وست چستر

مشارکت کنندگان:

جان هکر رایت
گروه فلسفه دانشگاه گوئلف

دنیل سینکلر
دانشکده حقوق دانشگاه فوردهام

پیتر دی هرشوک
مرکز شرقی-غربی

آندریا سالیوان کلارک
گروه فلسفه دانشگاه ویندزور

فصل ۱

مقدمه: ماشین‌های اخلاقی

ماشین‌های اخلاقی

این کتاب، برای دانشمندان داده و افراد علاقه‌مندی است که در برخی جهات آن دچار تردید شده‌اند، یا به طور خلاصه، راه درست و غلط استفاده از دیتا را به افراد نشان دهد و از استفاده‌ی غیراخلاقی آن جلوگیری کند.

به نظر می‌رسد که اهمیت علم داده در زندگی روزمره بسیار کم است! این امر باعث می‌شود که مردم عادی حتی در درک کردن این فناوری قدرتمند، کاملا ناتوان باشند؛ چه رسد به شکل‌دهی یا اداره‌ی آن! از طرفی خیلی از دانشمندانی که در این زمینه مشغول فعالیت هستند، نه زمان کافی برای کسب معلومات اخلاقی را دارند و نه منابع کافی برای اینکه ذهن خود را در گیر اهمیت اخلاق در این زمینه کنند. در صورتی که این فناوری می‌تواند تأثیرات اخلاقی زیادی را بر جامعه وارد کند. این کتاب در جهت کاهش این کمبودها نوشته‌شده است تا چراغ راهی باشد برای کسانی که به اخلاق در این حوزه اهمیت می‌دهند. برای این ایده که «دانشمندان باید اخلاق را بیاموزند»، تفکراتی مانند «شما نمی‌توانید چیزی در مورد اخلاق به کسی بیاموزید، مردم آن را میسازند» وجود دارد. البته قسمتی از آن درست است، مردم یک جامعه، اخلاق را می‌سازند.

امروزه ما با یک پدیده‌ی بسیار قدرتمند و البته بسیار پر خطر به نام «علم داده» روبرو هستیم؛ بنابراین، باید اخلاقیات و ضوابط این حوزه به صورت گسترده آموزش داده شود.

برای اینکه این کتاب تا حد امکان مفید و دوستانه واقع شود، سعی کردیم مطالب را با لحنی

ساده بیان کنیم. در این کتاب، ۷ مثال واقعی که استفاده نادرست از علم داده را نشان می‌دهند، بیان می‌کنیم. ما همچنین با چندین دانشمند برجسته‌ی اخلاق تماس گرفتیم تا در هر مورد نظراتشان را بپرسیم. همچنین برای ارائه‌ی طیف وسیع رفتارها و اخلاقیات انسانی، از سه دیدگاه غرب نسبت به اخلاق، فراتر رفتیم، سه رویکرد غرب عبارتند از: نتیجه‌گرایی(فایده‌گرایی)، دین شناسی و رفتار با تقوا. رویکردهایی که به طور اضافی بررسی کردیم: بودایی، یهودی، بومی و آفریقایی. هر یک از این رفتارها و رویکردها، می‌توانند زاویه‌ی دید متنوعی را ارائه کنند که ممکن است به آن فکر نکرده باشیم. هدف ما این است که درک کاملی از هر رویکرد ارائه دهیم، یک جعبه ابزار کامل برای روبرویی با چالش‌های آینده.

همانطور که می‌دانیم، یک مشکل خاص، می‌تواند با زوایای دید متفاوت (رویکردهای اخلاقی متفاوت که اشاره کردیم)، به طور مختلف تحلیل و بررسی شود. توانایی تحلیل معضل از دیدگاه‌های مختلف، لازمه‌ی «تفکر انتقادی» است. امیدوارم این کتاب دیدگاه گسترده‌ای را در اختیار خواننده قرار دهد!

علم داده چیست؟

علم داده، اصولی است برای استخراج دیتاهای غیر بدیهی و الگوها از مجموعه دیتاهای بزرگ. از طرفی هوش مصنوعی را می‌توان هر گونه پردازش اطلاعات که کارکرد روانی را انجام می‌دهد، اطلاق کرد. مثلا پیش‌بینی، تداعی کردن، تخیل کردن، برنامه‌ریزی و به طور کلی، هر پردازشی که تا کنون موجودات زنده قادر به انجام آن بودند.

ماشین لرنینگ ،*(ML)* زیرمجموعه‌ای از علم داده و بخش رو به رشدی از این زمینه است. بر خلاف ،*GOFAI* ماشین لرنینگ *(ML)* شکلی از هوش مصنوعی است که از رویکردهای آماری برای یافتن الگوها در دنیا (که بهم ریخته است) استفاده می‌کند. در خیلی جهات، *(ML)* پاسخی برای شکست‌های زودهنگام هوش مصنوعی سمبلیک *(GOFAI)* در بیرون از فضای آزمایشگاهی بود، به دلیل اینکه *GOFAI* قادر به پردازش پیچیدگی دنیای واقعی نبود.

الگوریتم‌های ،*ML* با لایه‌های موازی اطلاعاتی که ارائه می‌شوند، آموزش داده می‌شوند و می‌توانند به روش‌هایی بیاموزند که نظارت نشده و نسبتا مرموز هستند! خیلی شبیه عملکرد مغز ما (از یک روش یا تابع استفاده می‌کند و آن را بر روی مجموعه‌ای از دیتا اعمال می‌کند). مانند تابعی که ایمیل‌های به درد نخور (هرزنامه) را شناسایی می‌کند؛ این تابع بر روی مجموعه‌ای از ایمیل‌ها اعمال می‌شود یا مشخص شود که کدام ایمیل به درد نخور است.

ویژگی‌های هرزنامه‌ها و غیر هرزنامه‌ها قبلا توسط انسان‌هایی که تفاوت را می‌دانند، برای الگوریتم برچسب گذاری می‌شود. از طرف دیگر، یادگیری بدون نظارت، شامل هیچ برچسب‌زنی‌ای نمی‌شود و ما نمی‌دانیم که دنبال چه فاکتورهایی هستیم! این الگوریتم در ابتدا مجموعه‌ای دیتا دریافت می‌کند و بررسی می‌کند که کدام ویژگی‌ها مرتبط هستند. برای مثال، یک الگوریتم بدون نظارت، ممکن است که به تصاویر متعددی از سگ نگاه کند و تعیین کند چه ویژگی‌هایی جوهره‌ی «سگ بودن» را به وجود می‌آورد. زمانی هم که با یک تصویر جدید روبرو می‌شود، می‌تواند تصمیم بگیرد که سگ

است یا خیر.

امروزه ابزارهای علم داده خیلی کاربرپسندتر شدند و تازه‌واردان و حتی افرادی که آموزش کمی دارند، به راحتی می‌توانند وارد این زمینه شوند. این به این معنی است که هیچ وقت انجام کار با نتایج بد در این زمینه، به این آسانی نبوده است! بنابراین عواقب پروژه‌هایی بد، باید توسط کسانی که وظیفه‌ی طراحی یا اجرای آن را دارند، پیش‌بینی شود.

همانطور که کِلِهِر *(Kelleher)* توضیح می‌دهد: «دیتا یا داده»، عنصری است که از دنیای واقعی انتزاع شده است و «اطلاعات»، داده‌هایی هستند که سازماندهی شدند تا مفید واقع شوند و «دانش» درک دقیق اطلاعاتی هست که داده‌ها به ما می‌دهند. اما با ارزش‌تر از همه، خرد است؛ که زمانی رخ می‌دهد که دانش را برای هدف خوب به کار ببریم. هدف ما این است که به خوانندگان خود کمک کنیم تا این خرد را توسعه دهند؛ که فکر می‌کنیم در قلب اخلاق علم داده قرار دارد.

بنابراین، اخلاق فقط بخشی از انجام خوب علم داده است. این یعنی، یک مشکل در دنیای واقعی، بسیار فراتر از جنبه‌های فنی آن است و البته اینکه یک سیستم چگونه قرار است زندگی افراد را تحت تاثیر قرار دهد نیز، اهمیت دارد!

موارد مطالعاتی

مورد۱ – مورد اول اخلاق تحقیق و روش علمی

مورد مطالعاتی اول، خواننده را با مفاهیمی مانند تکثیرپذیری، دقت و اعتبار آشنا می‌کند. بسیاری از این بحث‌ها بر اساس تلاش‌های اخیر در روانشناسی و همچنین علوم اجتماعی و پزشکی استوار شده است تا به واقعیتی که قسمت قابل توجهی از نتایج منتشر شده قابل تکثیر یا اعتبارسنجی نیستند، پاسخ دهند.

این مورد، سوءرفتار تحقیقاتی در آزمایشگاه غذایی کورنل *(Cornell Food and Brand Lab)* به وسیله‌ی برایان وانسینک *(Brian Wansink)* را شرح می‌دهد. مشخص شد که او برای نتایج از چندین روش غیر علمی و البته غیر اخلاقی استفاده کرده است. روش‌هایی از جمله: *cherry picking* (علنی کردن نتایج دلخواه)، روش *HAEKing* (فرضیه سازی پس از مشخص شدن نتایج تجربی) و روش *p-hacking* (دستکاری داده ها برای به دست آوردن یک نتیجه آماری معنی دار).

آقایان «سینگر» و «فای تسه» تفسیری بر رفتار «وانیسنک» از دیدگاه فایده‌گرایی ارائه می‌دهند. این دو بر اهمیت راست بودن نتایج علمی که دیگران به آن تکیه می‌کنند، تأکید دارند. کسانی که این وظیفه را به عهده گرفته‌اند تا شواهد علمی و تجربی‌ای را که دیگران از آن استفاده می‌کنند، ارائه دهند، درواقع بار سنگینی را بر دوش دارند. آن‌ها باید این کار با به بهترین نحو ممکن انجام دهند.

مورد۲ - مدل‌های ماشین در دادگاه

الگوریتم‌های *ML*، در چندین پرونده‌ی جنایی مورد استفاده قرار گرفت و البته اشکال اخلاقی را نیز به بار آورد. حتی بهترین مدل‌های تائید شده نیز در زمینه‌های اجتماعی مختلف نیز عملکر متفاوتی دارند. حتی مدل‌های عالی نیز که توسط انسان استفاده می‌شوند، می‌توانند عواقب ناخواسته‌ای را شامل: «تبعیض»، «تعصب» و «سوءاستفاده‌ی عمدی» به بار بیاورند.

استفاده از مدل *Markov chain Monte Carlo (MCMC)* منجر به معضل اخلاقی می‌شود. زیرا این مدل نمی‌تواند به طور کامل تکرار شود و درنتیجه شواهد تولید شده توسط آن نیز قابل تکرار نیست.

این مسائل از طریق یک مطالعه درباره الگوریتم‌های ترکیب *DNA* و نقش آن‌ها در محاکمه نادرست «اورال نیکولاس هیلاری» *(Oral Nicholas Hillary)* در قتل یک پسر جوان در پاتسدام، نیویورک، توضیح داده می‌شوند. در این مورد، تحقیقات پلیس و محاکمه شامل **عوامل**

قویی از تعصب شخصی و نژادی علیه هیلاری بود، که یک مربی محبوب و موفق با ریشه‌های آفریقایی-کارائیبی بود. این موجب شد تفسیر بسیار تعصب‌آمیزی از شواهد DNA برای متهم، ساخته شود. بازرسی شد که شواهد ناقص و غیرقابل اعتماد بوده و به درستی توسط دادگاه از پرونده حذف شده است، که در نتیجه به هیلاری تبرئه شد.

«ساموئل جِی لووین» از دیدگاه اخلاق یهودی، استفاده از مدل‌های یادگیری ماشینی در سامانه عدالت کیفری را مورد بررسی قرار می‌دهد و به بررسی تنش‌ها (در واقع تناقض) بین جبر و اراده آزاد در اخلاق یهودی پرداخته است.

ما همه عاملان اخلاقی هستیم و مسئول انتخاب‌های خودمان هستیم. اما اگر اعمال ما پیش از این تعیین شده باشند، آیا ما در حقیقت دیگران را برای تصمیماتی که نگرفته‌اند، قضاوت نمی‌کنیم؟ طوری دیگر بیان می‌کنم: اگر اعمال ما از پیش تعیین شده باشند، این به معنای آن است که پیش‌تر مشخص شده‌اند و نشان می‌دهد که ما کنترلی بر روی آن‌ها نداریم. جمله‌ای که شما ارائه داده‌اید، سؤالی را مطرح می‌کند که آیا عادلانه است که افراد را برای تصمیماتی که آن‌ها جز گزینه‌هایی که می‌توانستند انتخاب کنند، قضاوت کنیم؟ به عبارت دیگر، اگر انتخاب‌های یک شخص از پیش تعیین شده باشد و او اراده آزادی نداشته باشد، آیا منصفانه است که او را برای آن تصمیمات مسئول دانسته و قضاوت کنیم؟

این را می‌توان در استفاده گسترده از مدل‌های ماشینی برای پیش‌بینی میزان تکرار جرم و احکام و تصمیم‌گیری برای دریافت وثیقه برای متهمان مشاهده کرد. بسیاری از قوانین جزایی ما مبتنی بر ایده‌هایی در مورد اختیار است که از یهودیت گرفته شده و از طریق ادیان ابراهیمی منتشر شده است. تنش بین جبرگرایی و اراده آزاد در بسیاری از تصمیم گیری ها در سیستم عدالت کیفری ما نفوذ می کند. در همین حال، کسانی که به‌جای عدالت به دنبال قدرت هستند، می‌توانند از فناوری‌های علمی به گونه‌ای سوء استفاده کنند که از مرزهای اخلاقی خارج شود.

«کالین مارشال» اثبات‌های تولیدشده توسط مدل‌های ماشین را از دیدگاه اخلاق «دانتولوژیک» بررسی می‌کند، که به مدت طولانی نگران شناسایی و از بین بردن اشکالاتی از نوع ناعادلانه‌گرایی در

تصمیم‌گیری اخلاقی بوده است که برخی افراد را نسبت به دیگران در مزیت قرار می‌دهد. تصمیمات اخلاقی باید آزمون همگانی را پشت سر بگذارند: اگر یک اقدام همگانی نباشد، به این معناست که همه انجام‌دهندگان در یک موقعیت مشابه، به احتمال زیاد یک انتخاب مشابه را انجام خواهند داد. این باید همه دانشمندان داده را تشویق کند که به تأثیرات مدل‌هایشان از دیدگاه افرادی که تحت تأثیر قرار می‌گیرند نگاهی بیندازند. این نیازمند این است که عاملان اخلاقی از دیدگاه خود، گاهی اوقات فایده‌گرایی، خارج شوند. اگر همه به این شیوه عمل کنند، سیستمی که ما می‌خواهیم در آن به طور غیرعادلانه توسط یک تحلیلگر جزئی یا یک الگوریتم تعصبی اتهام شویم، چگونه خواهد بود؟ سیستم‌هایی که شامل ناعدالتی غیرمشروع هستند، از نظر اخلاقی غیرمجاز هستند و باید استفاده نشوند.

مورد۳ - رسانه‌های ساختگی و خشونت سیاسی

در این مورد، دو مثال را بررسی می‌کنیم:

۱- مثال اول در مورد سخنرانی رئیس جمهور «گابن علی بونگو» *(Gabonese Ali Bongo)* در شب سال نو سال ۲۰۱۹ است. این ویدئو برای فرونشاندن ترس‌ها در مورد بیماری اخیر «بونگو» طراحی شده بود. اما زمانی که این ویدئو به عنوان یک دیپ‌فیک *(deepfake)* شناخته شده، تنش‌های سیاسی را برانگیخت. سربازان گارد جمهوری خواه، کودتای نافرجامی را در لیبرویل راه انداختند؛ به این دلیل که «بونگو» دیگر در رأس کار نیست و نمی‌توان به حزب حاکم اعتماد کرد. کودتا با خشونت سرکوب شد و منجر به کشته شدن دو سرباز و بازداشت خیلی‌ها شد. ولی بعدا مشخص شد که ویدئو کاملا واقعی بوده! در ویدئو به نظر می‌رسد که اثرات *(deepfake)* روی «بونگو» مشاهده می‌شود، ولی درواقع تاثیرات بعد از عمل باعث این موضوع شده بود! چشم‌های «بونگو» به طور غیر طبیعی در ویدئو مشاهده می‌شود که گمان *(deepfake)* را می‌رساند.

۲- مثال دوم به «فیک‌های کم عمق» *(shallow fakes)* علیه زنان در هند، به ویژه روزنامه نگاران و سیاستمدارانی که از حزب حاکم انتقاد می کنند، می پردازد. این رسانه‌های دستکاری شده شامل سایت‌های حراج جعلی هستند که مدعی «فروش» زنان هستند و آن‌ها را در شرایط تحقیرآمیز جنسی «پورنوگرافی» به تصویر می‌کشند. خبرنگاران زن در هند متوجه شده‌اند که پورنوگرافی تقلبی می‌تواند باعث ویرانی روحی‌شان شود و زندگی حرفه‌ای آن‌ها را به پایان برساند فقط به این دلیل که در دسترس عموم قرار گرفته‌اند! و نه به این دلیل که کسی باور کند این عکس‌ها و ویدئوها "واقعی" هستند.

در واقع، رسانه‌های مصنوعی در زمینه‌های بیشتری وارد زندگی ما می شوند. محتوایی که می‌خوانیم به‌طور فزاینده‌ای توسط هوش مصنوعی تولید می‌شود! پدیده‌ای که اخیراً حتی در مجلات علمی با داوری مشابه نیز دیده می‌شود. یعنی هوش مصنوعی مطلب علمی تولید می‌کند!

معنای زندگی در دنیای رسانه‌های دستکاری شده چیست؟ دنیایی که دیگر نمی توان حقیقت را به طور قابل اعتماد تعیین کرد و روی آن توافق کرد، یا حتی **در دنیایی که حقیقت دیگر اهمیتی ندارد؟** الگوریتم‌های هوش مصنوعی ممکن است به ما در شناسایی و حذف رسانه‌های مصنوعی کمک کنند، اما نمی‌توانند این مشکلات عمیق‌تر را برطرف کنند.

«سینگر» و «فای تسه» به مشکلات ناشی از رسانه‌های مصنوعی (شبکه‌های اجتماعی دستکاری‌شده) از دریچه اخلاق فایده گرایانه نگاه می کنند. آن‌ها بر اهمیت حقیقت تأکید می کنند که (حقیقت) به معنای استفاده از روش علمی و شواهد تجربی معتبر برای تصمیم گیری است. در غیر این صورت، فقط اعتمادمان را نسبت به نهادهای اصلی بیشتر از دست می‌دهیم. اعتماد و نهادهای قوی (اصلی) که رفاه افراد جامعه ما را ارتقا می دهند، با ارزش گذاری ما ساخته می شوند. پورنوگرافی چه به صورت کم عمق و چه به صورت عمیق، به طور ویژه سلامتی را تخریب می کند و توسط اخلاق «فایده گرایانه» رد می‌شود. این (پورنوگرافی عمیق و کم عمق) در خدمت تقویت این عقیده که: "زنان وسیله‌ای برای سرگرمی دیگران هستند" است. این امر می‌تواند آسیب‌هایی از جمله: ارعاب (ایجاد رعب و وحشت برای زنان)، ظلم و ستم و وادار کردن زنان به انجام کارهایی خلاف قوانین

زندگی عادی، داشته‌باشد.

«مورانگی» از دیدگاه اخلاق «اوبونتو» می‌نویسد. [به طور کلی، «اخلاق اوبونتو به عنوان مجموعه‌ای از ارزش‌ها تعریف می‌شود که از میان آن‌ها می‌توان به روابط متقابل، خیر مشترک، روابط مسالمت‌آمیز، تأکید بر کرامت انسانی، و ارزش زندگی انسانی و نیز اجماع، مدارا، و احترام متقابل اشاره کرد».] او تشویق می‌کند که داده‌شناسان نقش خود را به عنوان معماران جهانی که در آن زندگی می‌کنیم درک کنند و بر پیامدهای ساخت و ساز جهان خود تأمل کنند. او خاطرنشان می کند که علم داده در حال حاضر به عنوان یک تلاش خنثی و غیرسیاسی تدریس می‌شود، اما اثرات آن بر مردم آفریقا چیزی جز خنثی است. در اخلاق بومی آفریقایی، رفاه جامعه هم‌زمان، اخلاقی و سیاسی است و شامل هر دو «فرد» و «جامعه» می‌شود. هوش مصنوعی هر دو را با تجاوز به (فرهنگ) جوامع در آفریقا و سراسر جهان، و با تضعیف حس مشترک، نظم اجتماعی بومی و ارزش‌های اجتماعی که اساس زندگی‌های اصیل و اخلاقی را تشکیل می دهند، تضعیف می کند.

«میلر» و «سالیوان-کلارک» از اخلاق بومی برای بحث در مورد راه‌های مختلف استفاده از داده‌ها برای دستکاری، اجبار، کنترل، سرکوب و سلب حق رای گروه‌های خاص استفاده می‌کنند. افراد بومی اغلب هدف داده‌هایی با هدف مشخص، از این طریق بوده‌اند. این امر، منجر به رشد «جنبش حاکمیت داده‌های بومی» شده است که استقلال و کنترل بر داده‌هایشان و نحوه‌ی استفاده از آن‌ها را به خودشان برمی‌گرداند.

مورد۴ - بیومتریک و تشخصی چهره

در این مورد، به بررسی مسائل اخلاقی ناشی از استفاده از «بیومتریک» به عنوان نوعی کلیدشناسایی می‌پردازیم. در دنیایی که اطلاعات مانند گنج است، بسیار مهم است تا افراد ابزار معتبری برای احراز هویت و جلوگیری از دسترسی به دیتاهای خصوصی خود داشته‌باشند. کلیدهای (فرم‌های) بیومتریک، خیلی از این مشکلات را حل می‌کنند؛ آن‌ها کلیدهایی کامل و قابل اعتماد هستند و از

سایر اَشکال شناسه‌ها (پسورد، شماره‌ی تلفن، ایمیل، الگوها و ...)، امن‌تر هستند. هرچند، زمانی که بیومتریک‌ها توسط مقامات برای نظارت و پزشکی قانونی استفاده شوند، می‌توانند مشکلات اخلاقی ایجاد کنند. در این مورد، ما به استفاده‌ی نادرست «پلیس سواره‌ی سلطنتی کانادا *(RCMP)*» اشاره می‌کنیم، که توسط کمیسیونر حریم خصوصی کانادا؛ به عنوان نقض قوانین حریم خصوصی شناخته‌شد.

افراد *(RCMP)* از سیستم تولید شده توسط شرکتی به نام *(Clearview AI)* برای جستجوی مظنونان و یافتن کودکانِ قربانی استثمار جنسی آنلاین، استفاده کرده بود. عکس‌های استفاده شده توسط *Clearview*، از شبکه‌های اجتماعی و سایر سایت‌های اینترنتی گرفته شده بود و اسماً «عمومی» بود. بنابراین *RCMP* استدلال کرد که استفاده آن‌ها از این عکس‌ها قوانین حریم خصوصی را نقض نمی کند. ولی با این حال باید توجه داشت که دیتای گرفته‌شده از سایت‌های عمومی نیز باید با رضایت کابر آن دیتا باشد؛ وگرنه منجر به نقض مشکلات حریم خصوصی می‌شود.

«داودزول» و «گلتز» اظهار داشتند که تصمیم هوش مصنوعی *Clearview* برای به کارگیری فناوری تشخیص چهره در جنگ اوکراین، هم قوانین بین‌المللی درگیری‌های مسلحانه و هم ارزش‌های بشردوستانه را نقض می‌کند. ما استدلال می‌کنیم که سیستم‌های داده‌ای که پتانسیل هدف قرار دادن غیرنظامیان یا نقض قوانین درگیری مسلحانه را دارند، غیراخلاقی هستند و استفاده از آن‌ها باید ممنوع شود.

«میلر» و «سالیوان کلارک» داده‌های بیومتریک را از منظر ارزش‌های بومی تجزیه و تحلیل می کنند. داده‌های بیومتریک به خودی خود غیراخلاقی و مضر نیستند، اما ممکن است به روش‌هایی غیراخلاقی و آسیب رسان مورد استفاده قرار گیرند. این امر، به خوبی توسط افراد بومی درک می‌شود، زیرا اغلب تجربه کرده‌اند که از داده‌ها، علیه آن‌ها استفاده شده است. برای بررسی حریم خصوصی، نباید یک دیدگاه و زاویه‌ی دید را برای همه تعمیم داد، بلکه برای هر قوم یا گروه، باید ارزش‌ها و هنجارهای آن‌ها و حتی ارزش‌های بومی را نیز دخیل کرد. بررسی این موضوع به صورت تک بعدی، کار درستی نیست. می‌توانیم از ارزش‌های فطری برای این سؤال استفاده کنیم که آیا استفاده از

داده‌های بیومتریک، تعادل و هماهنگی در روابط را مختل می‌کند؟ آیا استفاده از داده‌های بیومتریک در دادرسی کیفری، باعث ایجاد حس اعتماد بیش‌ازحد به گناهکاری متهم می‌شود؟ و باعث می‌شود که از فروتنی که یک ارزش فطری است، دلسرد شود؟ راه درست این است که برای استفاده از دیتای مردم، از آن‌ها اجازه گرفته شود و البته حاکمیت و خودمختاری برای داده‌هایشان را به مردم برگردانده شود.

مورد۵ - تعدیل محتوا: سخنرانی خطرناک و پاکسازی قومی در میانمار

این مطالعه موردی به بررسی استفاده از هوش مصنوعی در تعدیل محتوا می پردازد یعنی اینکه چگونه باید محتوا و دیتا در فضای مجازی کنترل شود. مثالی هم از سخنرانی ضد «روهینگیا» (شهری در میانمار) در *Facebook* که پاکسازی قومی آن‌ها توسط نیروی دولتی در میانمار را در پی داشت. این موضوع، بحث‌های زیادی در مورد اینکه چه محتوایی باید در پلتفرم‌های شبکه‌های اجتماعی ممنوع شود ایجاد کرده‌است.

الگوریتم‌های *ML*، باید در کنار نیروی انسانی کار کنند تا مؤثر باشند. در عین حال، تعدیل و کنترل محتوای گذاشته‌شده توسط انسان، کاری سخت و خطرناک است؛ زیرا می‌تواند شکایت صاحبان محتوا را در پی داشته‌باشد.

شرکت‌های شبکه‌های اجتماعی، در حال توسعه‌ی دستورالعمل‌های تعدیل محتوا هستند و البته هیچ‌وقت معلوم نیست که چه محتوایی باید ممنوع شود! محدودیت‌هایی که توسط هوش مصنوعی شناسایی و حذف می‌شود، می‌تواند تاثیر دلخراشی در «آزادی بیان» داشته‌باشد. برای کسانی که دیتایشان حذف و کنترل شده، اغلب اطلاعات کمی در اختیار می‌گذارند و برای کسانی که آزادیشان محدود شده، هیچ توسلی وجود ندارد.

در عین حال، برای کسانی که از طریق تهدید، آزار و اذیت، پورنوگرافی جعلی، رادیکالیسم یا رادیکال‌سازی (هواداری از تغییرات ریشه‌ای در جامعه، تغییرات بنیادی و ریشه‌ای) یا کلاهبرداری

توسط محتوا در شبکه‌های اجتماعی آسیب دیده‌اند، اغلب راه‌حل‌های کمی در برابر پلتفرم‌های شبکه‌های اجتماعی قرار دارد.

«هرشاک»، محتوای مدیریت را از دیدگاه اخلاق بودایی تحلیل می‌کند. Facebook مسئولیت اخلاقی دارد که تعقیب منافع تجاری خود را به گونه‌ای انجام دهد که آسیب نرساند، و وقتی که اجازه داد تا سخنان بد و نفرت‌انگیز در برابر «روهینگیا» در پلتفرم خود گسترش یابد، این مسئولیت‌ها را نادیده گرفت. «هرشاک»، اشاره به ابهام مرزهای اخلاقی توسط پلتفرم‌هایی مانند *Facebook* دارد. این کار باعث پخش مسئولیت و آسیب در بین گروه‌ها، افراد و عوامل متنوع می‌شود.

در تصمیم‌گیری درباره نحوه‌ی مدیریت محتوا در آینده، ارزش‌های «بودایی» به ما یاد می‌دهند که از سوءاستفاده، داستان‌سازی و شایعه، غیبت، تهمت، دروغ و نفرت که در شبکه‌های اجتماعی بسیار شایع هستند، پرهیز می‌کنند. این ویژگی‌های اخلاقی و رفتاری در بودیسم (آئین بودایی) ارزشمند است: «شفقت، مهربانی، متانت، و شادی در اقبال دیگران» - که به وضوح وجود ندارند و ما می‌توانیم در تمام تعاملات خود با دیگران، از جمله در شبکه‌های اجتماعی، آن‌ها را پرورش دهیم.

«هکر-رایت» از دیدگاه اخلاق فضیلت‌محور به سخنان نفرت‌انگیز و بد در مدیریت محتوا نگاه می‌کند. چه نوع فضایلی باید از طریق پاسخ‌های ما به مدیریت محتوا ترویج یابند یا کنار گذاشته شوند؟ ما هرگز نمی‌توانیم همه محتوای مضر را از شبکه‌های اجتماعی حذف کنیم، و این به تنهایی باعث ارتقای یک جامعه نمی‌شود. از این گذشته، رسانه‌ها دقیقاً به این دلیل مؤثر هستند که ما به آن‌ها اجازه می دهیم دیدگاه‌های قبلیِ ما را تقویت کنند. به این ترتیب، همه ما خواسته یا ناخواسته، در دستکاری رسانه‌های اجتماعی شرکت می‌کنیم (منظور این است که در رسانه‌ها، خواسته یا ناخواسته، فعالیت می‌کنیم). ما می‌توانیم با پرورش فضیلت‌های مهمی مانند شجاعت اخلاقی، تفکر انتقادی و تمایل به ارتقای جامعه (از جمله کسانی که با ما مخالف هستند) با این کار مقابله کنیم. به این ترتیب، می‌توانیم به نوعی خرد عملی دست یابیم که از پرورش عادتی و صد البته آگاهانه‌یِ فضایلی که ارسطو معتقد بود منجر به رفاه است، ناشی می‌شود.

«میلر» و «سالیوان کلارک» به تعدیل محتوا از دریچه اخلاق فطری و ذاتی نگاه می‌کنند که

همه چیز را به هم مرتبط می بیند. *Facebook* درک نکرد که محتوای موجود در پلتفرم آن‌ها چگونه روی «روهینگیا» تأثیر می‌گذارد. الگوریتم‌های آن‌ها، سخنرانی‌های بسیار جذاب را در اولویت قرار می‌داد، حتی زمانی که نفرت و دشمنی را ترویج می‌داد و خشونت علیه یک گروه آسیب‌پذیر را تقویت می‌کرد! این اقدامات باعث ایجاد نوعی ناهماهنگی و عدم تعادل می‌شود که اغلب باعث آسیب می‌شود. *Facebook* همچنین نتوانست ارزش مهم فروتنی در تفکر را پرورش دهد. کسانی که سیستم‌های *ML* را طراحی و اجرا می‌کنند، موظفند از محدودیت‌های الگوریتم‌های خود و همچنین پتانسیل آن‌ها برای سوءاستفاده، آگاه باشند. «میلر» و «سالیوان-کلارک» به نکته‌ی مهمی اشاره می‌کنند که چندین مشارکت‌کننده به آن اذعان دارند، این که: «کلمات قدرت دارند». الگوریتم‌هایی که گفتار خاصی را برای دیگران تبلیغ می‌کنند یا باعث ایجاد ابهام می‌شوند نیز قدرت دارند، و باید با فروتنی و در نظر گرفتن رفاه دیگران از آن‌ها استفاده کرد.

مورد۶ - بدافزار ذهنی: الگوریتم‌ها و معماری انتخاب

در سال 2013، شرکت تجزیه و تحلیل داده «کمبریج آنالیتیکا» شروع به جمع‌آوری اطلاعات در *Facebook* برای ایجاد پروفایل‌های روانشناختی عمیق روی ده‌ها میلیون کاربر بدون، رضایت آن‌ها کرد. سپس این داده‌ها به بازاریابان فروخته شد، از جمله چندین کمپین سیاسی. این رسوایی منجر به ورشکستگی «کمبریج آنالیتیکا» و میلیاردها دلار جریمه برای *Facebook* شد! رسوایی «کمبریج آنالیتیکا» نشان داد که جمع‌آوری اطلاعات حساس روانشناختی از کاربران رسانه‌های اجتماعی و استفاده از این داده‌ها به روش‌هایی که آن‌ها را دستکاری می‌کنند (اغلب بر خلاف منافع مردم)، چقدر آسان است! این مورد، دقیقا یک مورد برجسته از چیزی است که ما «برافزار ذهنی» می‌نامیم. «بدافزار ذهنی» اغلب بر علیه کاربران به شیوه‌هایی استفاده می‌شود که نه صرفاً برای پیش‌بینی رفتار آن‌ها، بلکه برای تغییر رفتار آن‌ها طراحی شده‌اند (برای "تحت فشار دادن"، دستکاری و تغییر رفتارهای فردی و افکار عمومی).

بهره بردن از قدرت روانی الگوریتم‌ها برای افراد سیاسی آسان است. از زمان رسوایی «کمبریج آنالیتیکا»، انتقادات بیشتری مبنی بر: اینکه شرکت‌های شبکه‌های اجتماعی در مقابله با «لایک‌ها» و «فالوورها» نادرست و دستکاری شده و دیگر اشکال تعامل مصنوعی و غیرواقعی شکست خورده‌اند؛ و اینکه از این موضوع برای دستکاری در انتخابات و سرکوب استفاده شده است، وارد شده است.

«هرشاک» به تعدیل محتوا از دیدگاه اخلاق بودایی نگاه می کند. همیشه، همه‌ی شرکت‌های شبکه‌های اجتماعی، محتوا را برای کاربران خود فیلتر و تعدیل می‌کنند، و ما باید مراقب باشیم که آن‌ها چگونه این انتخاب‌ها را انجام می‌دهند، چه کسی مسئول این انتخاب است و چه ارزش‌هایی در اولویت هستند؟ معماری انتخاب دیجیتالی که ایجاد می‌کنیم باید رفاه فردی و اجتماعی را افزایش دهد. در حالی که نیاز به تقویت آزادی شخصی وجود دارد، ما باید آگاه باشیم که این پتانسیل این را نیز دارد که کاربران را در انتخاب‌های گذشته خود قفل کند و در نتیجه آزادی آن‌ها را محدودتر کند. این ممکن است به سادگی منجر به این شود که توده‌های بشریت «زندگی‌هایی را داشته باشند که در آن هرگز لازم نیست از اشتباهات درس بگیریم یا در رفتار سازگارانه شرکت کنیم». در اخلاق بودایی، همه چیز به هم مرتبط است. زیرساخت‌های دیجیتالی‌ای که ما ایجاد می‌کنیم نه تنها بر انتخاب‌ها، رفتار و روابط اجتماعی ما تأثیر می‌گذارد، بلکه اساساً چیزی که هستیم را تغییر می‌دهد!

«هکر رایت» بحث خود را در مورد انتخاب‌ها، به وسیله‌ی رفتار با فضیلت ادامه می‌دهد. «ارسطو» فاعل نیکوکار را کسی توصیف می‌کند که به دنبال خیر است و از منکر دوری می‌کند. یک مامور پاکدامن به دنبال خیر خواهد بود، اما آن‌ها همچنان به سمت برخی از رذایل کشیده می‌شوند و با انتخاب درست مبارزه خواهند کرد. یک مامور رذیل نیز با جذب خود به سمت رذیله، بدی و تباهی مبارزه می کند، اما آن‌ها قدرت اراده کافی را برای مبارزه ندارند. پس از شکست خوردن در مبارزه بین فضیلت و رذیلت، ممکن است که احساس شرمندگی کنند. از سوی دیگر ، یک ایده‌آل غلط (که فکر می‌کنند خوب و درست است) را پذیرفته اند، و بنابراین آن‌ها با تسلیم شدن به رذیلت، فکر می‌کنند که رذیلت برای زندگی خوب است، ولی در اشتباه‌اند.

شبکه‌های اجتماعی و شرکت‌های بازی‌سازی طراحی شده‌اند تا از طریق تاکتیک‌های هوشمندانه‌ی

دستکاری و غلبه بر قدرت اراده کاربران، همه را، به جز فضیلت‌ترین کاربران جذب کنند (فقط کاربران فضیل که درست‌کاراند، جذب نمی‌شوند). پرورش فضایل و تقویت اراده، می‌تواند ابزار موثری برای غلبه بر انبوه بدافزارهای ذهنی‌ای باشد که هر روزه با آن مواجه هستیم.

«مارشال» به بدافزارهای ذهنی از زاویه‌ی دئونتولوژیک (نگاه دینی) نگاه می‌کند. هرگونه تلاش برای تأثیرگذاری بر دیگران، ابتدا باید در جهت یک هدف اخلاقی باشد. در اخلاق دین شناسی، استفاده از دیگران به عنوان وسیله‌ای برای رسیدن به هدف ممنوع است. ما نمی‌توانیم برای تحقق منافع خود دیگران را زیر پا بگذاریم، کاری که بسیاری از شرکت‌ها و بازاریابان شبکه‌های اجتماعی انجام می‌دهند! دوم، هر نوع نفوذ اخلاقی و تأثیرگذاری باید مبتنی بر صداقت و گفتمان عقلانی و منطقی باشد. بدافزار ذهنی به دنبال جذب چیزی است که «کانمن» *(Kahneman)* آن را تفکر «**سیستم** 1» می‌نامد (تعریف سیستم 1: پاسخ‌های احساسی و خودکار (سریع و آسان) که در ابتدا به اطلاعات جدید می‌دهیم). با این حال، هر تلاش برای تأثیرگذاری، حتما باید روش تفکر «**سیستم** 2» را نیز درگیر کند (تعریف سیستم 2: روش‌های تفکر آگاهانه، منطقی و مشورتی (آهسته و دشوار)). کسانی که الگوریتم‌ها را به صورت اخلاقی و برای تأثیرگذاری به کار می‌برند، باید در مورد نحوه عملکرد الگوریتم‌ها صادق و شفاف باشند. درنهایت، ما باید فرآیند مشورتی منطقی تصمیم‌گیری بر اساس اطلاعات خوب و داده‌های تجربی صحیح را نسبت به پاسخ‌های سریع و احساسی اولویت دهیم؛ چیزی که امروز دقیقا برعکس آن در شبکه‌های اجتماعی درحال انجام است!

مورد۷ – هوش مصنوعی و موجودات غیر انسان

انسان‌ها تنها موجودات زنده‌ای نیستند که سیستم‌های *ML* بر منافعشان تأثیر می‌گذارند (چه به صورت مثبت و چه به صورت منفی). در این فصل، «سینگر» و «تسه» تحقیقات خود را در مورد روش‌هایی که الگوریتم‌های هوش مصنوعی بر رفاه حیوانات تأثیر می‌گذارند ارائه می‌کنند. اول، آن‌ها درباره‌ی تأثیرات مختلفی که نتایج موتورهای جستجو و الگوریتم‌های توصیه‌ای می‌توانند بر نحوه

تفکر ما در مورد حیوانات و در نتیجه نحوه برخورد ما با حیوانات تأثیر بگذارند، بحث می‌کنند. تعصب الگوریتمی در نتایج موتورهای جستجو و توصیه‌ی محتوا می‌تواند محتوا و تبلیغاتی را به ما ارائه دهد، که بر میزان تأثیر آن می‌افزاید. محصولات حیوانی‌ای که مصرف می‌کنیم در حالی که ظلم و آزار حیوانات در دنیای واقعی را پنهان می‌کنیم و کاربران را نسبت به این آسیب ها حساسیت زدایی می‌کنیم. مدل‌های زبانی، می‌توانند «بار نزادپرستی» زبان را تقویت کنند که حیوانات را تحقیر می‌کند. این تأثیر زیادی بر رفاه حیوانات دارد (منظور این است که فرضا صفت درنده برای ببر درست است، ولی درواقع یک صفت منفی به حساب می‌آید، مدل‌های زبانی ممکن است از این صفات استفاده کرده و ناخواسته محتوایی تولید کنند که گونه‌گرایانه و یا نژادپرستی را می‌رساند).

دوم، آن‌ها در مورد استفاده از هوش مصنوعی در مزارع و کارخانه‌ها بحث می کنند. مدل‌های ML در صنعت مزرعه‌ی کارخانه‌ای، برای جمع‌آوری اطلاعات در مورد حیوانات پرورشی، به منظور به دست آوردن سود حداکثری، استفاده می‌شوند. بیماری و مرگ و میر چقدر سود را به حداکثر می رساند؟ چه مقدار باید حیوانات تغذیه شوند تا رشد را، با پایین نگه داشتن هزینه‌ها متعادل کند؟ آن‌ها همچنین این موضوع مهم را مطرح می‌کنند که: چگونه رفتار حیوانات و حالات ذهنیشان را شناسایی و تفسیر می‌کنیم، زمانی که از دریچه چشم‌انداز خودمان، انسانی، نگاه می‌کنیم؛ ولی به راحتی حقوق‌شان را زیر پا می‌گذاریم. خروج هوش مصنوعی از ذهنیت انسانی و اتخاذ مجموعه‌ای از ارزش‌ها و دیدگاه‌های غیرانسانی به چه معناست؟ رفاه آینده حیوانات به نحوه حل این مسائل اخلاقی بستگی دارد.

«سینکلر» یک دیدگاه اخلاقی یهودی در مورد وظیفه رفتار با حیوانات به روشی درست و اخلاقی ارائه می دهد. در حالی که انسان‌ها نسبت به سایر موجودات برتری دارند، اولین مردم گیاهخوار بودند و بعدها که فاسد شدند، اجازه یافتند گوشت بخورند. مفهوم جلوگیری از ظلم به حیوانات عمیقاً در اخلاق یهودی گنجانده شده است، از جمله اجازه دادن به حیوانات کار برای استراحت در شب‌ها و لذت بردن از اوقات فراغت خود. «مارشال» درباره‌ی وضعیت اخلاقی حیوانات در اخلاق دئونتولوژیک بحث می کند. همه نسخه‌های اخلاق افضل (اخلاق وظیفه‌شناس، علماالاخلاق) اهمیت حقوق حیوانات

را به رسمیت می‌شناسند، گرچه در مورد اهمیت حقوق حیوانات، موارد متفاوت و استثنا هم وجود دارد. اگر چنین است، پس استفاده از حیوانات به عنوان ابزاری صِرف برای اهداف خود از نظر اخلاقی غیرمجاز خواهد بود. او همچنین به این نکته مهم اشاره می کند که عدم احترام کافی به ادعاهای اخلاقی و حقوق حیوانات می تواند باعث شود که به طور کلی به ادعاهای اخلاقی و حقوق دیگران نیز احترام نگذاریم. باید از بی‌تفاوتی نسبت به رنج و احساس دیگران به شدت اجتناب شود (هرکس با هر احساسی).

«مورانگی» تفسیری درباره حقوق اخلاقی حیوانات از منظر اخلاق آفریقایی ارائه می دهد. او نقشی را که استعمارزدایی در اخلاق هوش مصنوعی بازی می‌کند، بررسی می‌کند و اینکه آیا می‌توانیم اخلاق هوش مصنوعی را طوری توسعه دهیم که به دنبال درک ماهیت اشتراکی «ما» باشد که قلب رفتار «اوبونتو» را شرح می‌دهد؟ معماران این فناوری‌ها اغلب نمی‌توانند «خود را فرزندان هوش مصنوعی یا مادران و پدران هوش مصنوعی ببینند». درعوض، آن‌ها باید تشویق شوند تا به این فکر کنند که یک عامل اخلاقی به چه معناست، و چه چیزی به معنای رفاه است. این فضایی را برای یک هوش مصنوعی رهایی‌بخش به جای ظالمانه باز می‌کند (هوش مصنوعی‌ای که رفاه حیوانات را نیز ارتقا می‌دهد؛ زیرا اگر به اندازه‌ی کافی به اینکه چه کسی هستیم و چیستیم فکر نکنیم، نمی‌توانیم حیوانات را به عنوان موجوداتی که از حقوق برخوردار هستند، تصور کنیم).

همه‌ی مفسران این کتاب راهی به جلو برای دانشمندان داده ارائه می‌دهند تا در طراحی و استفاده از داده‌ها و سیستم‌های هوش مصنوعی اخلاق را رعایت کنند. در واقع، درگیر شدن با نظرات مشارکت‌کنندگان مطمئناً نوعی خرد را تقویت می‌کند که «کلهر» از آن حمایت کرده است، و این امر کمک زیادی به حرکت در دنیای الگوریتم‌ها و هوش مصنوعی می‌کند. اهمیت داده ها در این عصر قابل انکار نیست! این امر قدرت بزرگی را در دست دانشمندان داده قرار می‌دهد و همانطور که ضرب المثل قدیمی می گوید: "هرکه بامش بیش، برفش بیشتر". ما واقعاً امیدواریم که این کتاب ابزارهای ارزشمندی را ارائه دهد که به همه‌ی ما در انجام این مسئولیت بزرگ با عقل، شفقت و خرد کمک کند.

فصل ۲

مقدمه ای بر رویکردهای اخلاقی در علم داده

دانش علم فیزیک مرا به خاطر ناآگاهی از اخلاق، تسلی نمی‌دهد، اما علم اخلاق همیشه مرا به خاطر ناآگاهی از علم فیزیکی تسلی می‌دهد. (منظور نویسنده، تأکید مهم بودن علم اخلاق و ارزش‌های انسانی است)

Blaise Pascal 1624-1624

مقدمه

فناوری‌های یادگیری ماشین، در حال نفوذ به زندگی مردم عادی در سراسر جهان هستند. کاربران این فناوری‌ها، خواسته یا ناخواسته در زندگی اصولی دارند که رویکردهای آن، در میان فلفسه‌های غربی ارائه نشده. بنابراین، ما چندین رویکرد اخلاقی غیر غربی را در کتاب آورده‌ایم. این‌ها برای طراحان ارزش دانستن دارد، هم برای اینکه بتوانند کاوش اخلاقی خود را عمیق‌تر کنند و هم به این ترتیب که بتوانند بهتر درک کنند که چگونه فن آوری‌هایشان تفسیر، اتخاذ، استفاده و تنظیم می‌شود. ما خوش‌شانس بوده‌ایم که تفسیرهایی از دانشمندان برجسته در زمینه‌های اخلاق دئونتولوژیک، اخلاق نتیجه‌گرا (فایده گرا)، و اخلاق فضیلت و فطری، و همچنین از اخلاق اوبونتو، اخلاق بودایی، اخلاق یهودی، و اخلاق بومی و ذاتی دریافت کرده‌ایم. ما امیدواریم که این به خواننده دید وسیع‌تری بدهد تا درباره‌ی فناوری‌های یادگیری ماشین از دیدگاه‌های مختلف فکر کند و بفهمد که چگونه آن‌ها توسط جوامع سراسر جهان پذیرفته می‌شوند و چگونه عمل می‌کنند. هر یک از این رویکردهای اخلاقی در

زیر به اختصار آورده شده است.

رفتار نتیجه گرایی و فایده‌گرایی

توسط پیتر سینگر و ییپ فای تسه

نتیجه‌گرایی خانواده‌ای از نظریه‌ها است که بر این عقیده هستند که درست یا نادرست بودن یک عمل بستگی به پیامدهای آن دارد یا به عبارت دیگر، وضعیتی که اعمال باعث ایجاد آن می‌شود. فایده‌گرایی، در شکل کلاسیک خود، نظریه نتیجه گرایی است که منحصراً بر درد و لذت، یا شادی و بدبختی، به عنوان تنها پیامدهای اخلاقی مرتبط برای تعیین چگونگی ارزیابی پیامدهای اعمال تمرکز می کند. در اینجا تأکید بر این نکته حائز اهمیت است که فایده‌گرایی تنها در مورد ارزیابی درستی یا نادرستی اعمال نیست، بلکه در مورد ارزیابی خوب و بد حالت‌های امور است، که بی طرفانه در نظر گرفته می‌شوند. به طور خاص، فایده‌گرایان معتقدند که همه‌ی موجودات ذی‌شعور (آن‌هایی که می‌توانند درد و لذت را تجربه کنند) باید در نظر گرفته‌شوند و به علایق مشابه آن‌ها باید وزن مشابهی داده شود. در کنار هم، فایده‌گرایی، این دیدگاه است که یک عمل نه تنها باید منفعت برساند، بلکه از نظر اخلاقی نیز لازم است که بیشترین مازاد خالص ممکن را از شادی نسبت به بدبختی (یا لذت بر درد) به همراه داشته باشد. و هر عملی که بر خلاف این اصل باشد، ممنوع و غیرمجاز است.

اعتراضات رایج به سودگرایی

یک اعتراض رایج به سودگرایی این است که ما را به انجام اعمال آشکاراً غیراخلاقی هدایت می‌کند! «داستایوفسکی» در «برادران کارامازوف»، «ایوان» را به چالش می کشدکه یک نوزاد را تا سرحد مرگ شکنجه کند تا برای همه‌ی بشریت خوشبختی بیاورد. چالش «ایوان» به یک اعتراض معروف به سودگرایی تبدیل شده است. بیان ساختار اعتراض «داستایوفسکی» به طور رسمی این موضوع را

بهتر نشان می دهد:

فرض 1. اگر فایده‌گرایی درست بود، به درستی به ما می گفت که کدام اعمال درست و کدام نادرست است.

فرض 2. فایده‌گرایی به ما می گوید که اگر شکنجه‌ی یک کودک بی گناه تا حد مرگ عواقب بهتری نسبت به هر عمل دیگری به همراه داشته باشد، آنگاه شکنجه یک کودک بی گناه تا حد مرگ کار درستی خواهد بود.

فرض 3. شکنجه یک کودک بی گناه تا حد مرگ همیشه اشتباه است. نتیجه: فایده‌گرایی نادرست است.

بسیاری از ایرادات به فایده‌گرایی نیز به همین ترتیب مطرح می شوند: یک جراح به این فکر می‌کند که آیا مخفیانه اطمینان حاصل کند که یک عمل شکست می خورد؛ تا بیمار بمیرد و سپس از اعضای بدن او برای نجات جان چهار بیمار در انتظار اهدای اعضای ضروری استفاده شود. چنین نمونه‌هایی منعکس کننده‌ی دانش ما از نحوه عملکرد جهان نیستند. «ایوان» توضیح نداد که چگونه شکنجه‌ی کودک باعث شادی پایدار برای دیگران می‌شود. مثال پیوند عضو در نظر نمی‌گیرد که اگر کاری که جراح انجام داده مشخص شود، ممکن است منجر به عواقبی شود که بسیار بیشتر از مزایای مورد نظر است (ممکن است افراد نسبت به پزشکان بی‌اعتماد شوند). چگونه جراح می تواند کاملاً مطمئن باشد که او گرفتار نخواهد شد؟ این فرض که شکنجه یک کودک بی گناه همیشه اشتباه است، متکی به ذات و فطرت انسانی دارد. بنابراین وقتی با نمونه‌های عجیب و خیالی سروکار داریم، فرض 3 مشکوک است و نمی‌توان به آن به عنوان مبنایی برای رد فایده‌گرایی اعتماد کرد.

ایراد اصلی دیگر این است که اندازه‌گیری درد و لذت، یا شادی و غم است. سودگرایان سه پاسخ

اصلی به این اعتراض دارند. اولاً، این مشکلی محدود به فایده‌گرایی نیست. هر نظریه‌ی اخلاقی‌ای که مقداری به رفاه اهمیت می‌دهد از دشواری اندازه‌گیری رفاه افرادی که تحت تأثیر اعمال هستند نیز رنج می برد؛ و البته نظریه‌ی اخلاقی‌ای که تمام این ملاحظات رفاهی را نادیده می‌گیرد بسیار غیرقابل‌قبول خواهد بود.

ثانیاً، اگرچه اندازه‌گیری دقیق درد و لذت دشوار است، ترجیحات افراد و تا حدی حیوانات را می توان مشاهده، آزمایش و رتبه بندی کرد تا اولویت‌های آن‌ها آشکار مشخص. در برخی از مطالعات، روانشناسان با پرداخت هزینه به آزمایش‌شوندگان، سطوح خاصی از درد یا تحمل را در آن‌ها می‌سنجند. این موارد، اگرچه آن چیزی نیست که فایده‌گرایان کلاسیک آن را خیر می‌دانند، با این وجود، معیارهای مفیدی هستند که به ما ایده‌ای درباره‌ی درد و لذت می‌دهند. مدل دیگری که از موارد آشکار استفاده می‌کند، سال زندگی تعدیل‌شده با کیفیت یا *(QALY)*، حول این ایده است که یک سال زندگی با عملکرد یا سلامت مختل، به اندازه یک سال در سلامت عادی، خوب نیست. برای مثال، محققان از مردم می‌خواهند که خود را با آسیب‌های مختلف در سلامت تصور کنند (گاهی اوقات خود درد)، و سپس از آن‌ها می‌پرسند که حاضرید چند سال از زندگی خود را رها کنید تا این اختلال درمان شود؟ این روش اکنون در سطح جهانی توسط اقتصاددانان سلامت، محققان پزشکی و سیاست‌گذاران استفاده می‌شود. در نهایت، در اکثر موارد، عمل درست حتی بدون اندازه‌گیری واضح است. به عنوان مثال، پزشکی که ترتیب درد بیماران را در اولویت قرار می‌دهد، می تواند به وضوح ببیند که یک بیمار سوختگی، شدیدتر از فردی که از سرماخوردگی رنج می برد، درد دارد و در معرض خسر مرگ بسیار بالاتری است؛ بنابراین، باید بیمار سوختگی را در اولیت قرار داد. یا مثلا اگر فردی از شما بپرسد که نزدیک‌ترین رستوران گیاه‌خواران کجاست؟ شما به احتمال خیلی زیاد با ارائه‌ی اطلاعات درست، او را راهنمایی می‌کنید، تا اینکه اصلا جواب ندهید یا پاسخ اشتباه بدهید!

اگرچه مواردی هم وجود دارند که پس از تجزیه و تحلیل هم شفاف نیستند؛ ولی با این جود می‌توان تصمیمات معقولی گرفت. نکته‌ی مهمی که در اینجا باید مورد توجه قرار گیرد، این است

که نه تنها می توان بخش قابل توجهی از تصمیمات تحت فایده‌گرایی را بدون اندازه‌گیری لذت و درد اتخاذ کرد، بلکه آنچه در این دنیا در خطر است نیز معمولاً می‌تواند بدون اندازه‌گیری مستقیم درد و لذت تعیین شود. فقر جهانی (که باعث گرسنگی، تشنگی، بیماری‌ها و ... می‌شوند)، کشاورزی کارخانه‌ای و بیماری‌های همه‌گیر نمونه‌های مناسبی از مسائلی هستند که بدون شک، رنج عظیمی را برای تعداد زیادی از افراد به بار می‌آورند.

توصیه‌هایی برای به کارگیری صحیح اصول سودمندی

گسترده تر و طولانی تر فکر کنید

ما با «جان استوارت میل»، یک فایده‌گرای اولیه، موافقیم که باید «مفید بودن را به‌عنوان اصل نهایی در همه مسائل اخلاقی در نظر بگیریم. اما باید در فراگیرترین معنای آن فایده باشد». منظور ما از "فراگیرترین" این است که همه پیامدهای مرتبط، صرف نظر از زمان، فاصله فیزیکی، خویشاوندی و سایر ویژگی‌های اخلاقی نامربوط مانند جنسیت، نژاد، و عضویت در گونه باید در نظر گرفته شوند.

مسلماً، زمان یکی از بحث برانگیزترین آن‌هاست که از نظر اخلاقی نامربوط اعلام می‌شود. تخفیف زمان اغلب در زمینه‌های اقتصاد و یادگیری ماشین آموزش داده‌می‌شود و به کار می‌رود، ولی تصورات آن‌ها در مورد ترجیحات زمانی با ایده‌های فایده‌گرایی متفاوت است. در اقتصاد، کاهش زمان، برای دریافت لذت و خوشی در زمان کمتر مد نظر است؛ یعنی ما مایل هستیم که لذت و خوشی را در زمان کمتری به دست بیاوریم تا اینکه بخواهیم برای آن صبر کنیم. در یادگیری ماشین به ویژه یادگیری تقویتی، "ضریب تخفیف" (γ)، متغیری است که تعیین می‌کند که عامل، تمایل به اهداف و پاداش‌های زودهنگام دارد یا دیرهنگام (اهمیت را برای پاداش‌های فوری یا آینده تعیین می‌کند). اگر مقدار (γ) نزدیک به 1 باشد، عامل به پاداش‌های آینده، بیشتر اهمیت می‌دهد و در نتیجه تمایل دارد تا مسیری را که باعث رسیدن به هدف در آینده می‌شود، دنبال کند. به عبارتی، عامل تمایل دارد پاداش‌های آینده را بیشتر به صورت بلندمدت مد نظر قرار دهد. از سوی دیگر، اگر مقدار (γ)

نزدیک به 0 باشد، عامل بیشتر روی پاداش‌های فوری تمرکز می‌کند و تمایل دارد که از پاداش‌های فوری بهره‌برداری کند. به عبارتی، عامل در تصمیم‌گیری خود بیشتر به جوانب کوتاه‌مدت توجه می‌کند و پاداش‌های آینده، اهمیت نمی‌دهد. مثلا می‌توانیم بگوئیم به دلیل اینکه در آینده فلان بازار هدف وجود نخواهد داشت، (γ) را نزدیک به 0 د نظر می‌گیریم تا در کوتاه مدت، به نتیجه‌ی دلخواه برسیم، عکس این قضیه هم صادق است. به عنوان مثال، شکنجه در 100 سال به همان اندازه بد است که شکنجه‌ای اکنون به همان اندازه درد داشته‌باشد، اما اگر قطعیت کمتری داشته باشد (یعنی ممکن باشد که شکنجه انجام نشود)، ممکن است به همین دلیل آن را کاهش دهیم (یعنی شکنجه‌ی چیزی را می‌پذیریم که مارا شکنجه نکند یا آن موردی که قطعیت کمتری دارد).

بیایید سعی کنیم این اصول را در هوش مصنوعی و علم داده اعمال کنیم. برای مثال، در تصمیم‌گیری برای راه‌اندازی یک محصول، نه تنها باید تاثیری که ممکن است بر روی کاربران آن داشته باشد، بلکه باید در نظر داشت که چگونه جامعه وسیع‌تر افراد (در برخی موارد، حتی حیوانات) چه در کوتاه مدت و چه در بلند مدت ممکن است تحت تاثیر قرار گیرند. سؤالاتی از این قبیل باید پرسیده شود: آیا این محصول سوگیری‌ها، فرهنگ، ایدئولوژی‌ها، فضیلت‌ها یا سایر ارزش‌ها را در جامعه جذب و در نتیجه آن را تقویت می کند؟ آیا این محصول، یک صنعت بسیار ارزشمند را از بین می برد یا باعث به تاخیر انداختن یا جلوگیری از حذف یک صنعت غیراخلاقی می‌شود؟

از ارزش‌های مورد انتظار برای تصمیم گیری استفاده کنید

استفاده از تئوری ارزش مورد انتظار در تصمیم‌گیری، در تئوری تصمیم گیری، اقتصاد و علم داده، اساسی است (یعنی قبل از تصمیم‌گیری بسنجیم ببینم که دنبال چه چیزی هستیم و بر اساس آن تصمیم‌گیری بکنیم). اما باید در مورد نظریه‌های اخلاقی، به‌ویژه به حداکثر رساندن فاکتورهای اخلاقی مانند فایده‌گرایی نیز اعمال شود. مثال جراح در بخش قبل نشان می‌دهد که چرا سناریوهایی با ریسک بالا و کم احتمال، اهمیت دارند. مهم نیست که جراح چقدر با دقت سعی کرد عمل او را

مخفی نگه دارد، او نتوانست به طور منطقی به این نتیجه برسد که احتمال افشای راز صفر است. با توجه به اثرات مشخص کشف شدن راز (اگر کشف می‌شد، مردم نسبت به پزشکان اعتمادشان را از دست می‌دادند)، جراح باید به این نتیجه برسد که انجام چنین عملی اشتباه است.

در حالی که محاسبه ارزش مورد انتظار اغلب ساده است (انجام آن عمل، چندین انسان را نجات می‌داد)، به دلیل سوگیری‌های شناختی انسان (مثلا اینکه شما بیمار من رو به خاطر اهدای عضو، به عمد به قتل رساندید!)، اغلب به درستی استفاده نمی‌شود یا حتی اصلاً اعمال نمی‌شود. «غفلت احتمالی» یک سوگیری شناختی است که افراد نسبت به «عدم قطعیت‌ها» نشان می‌دهند، به‌ویژه «احتمالات کوچک»، که تمایل دارند یا به طور کامل از آنها غفلت کنند، یا تا حد زیادی (اغراق) آن را بزرگ کنند. یک مطالعه با دریافت اینکه مردم برای کاهش خطرات "رویدادهای نادر و پر تاثیر" یا ارزش خیلی زیاد یا بسیار پایین قائل هستند؛ (غفلت احتمالی) را تائید کرد. ما نیازی به جستجوی شواهدی مبنی بر غفلت جمعی از «رویدادهای نادر و پرتأثیر» نداریم! اگر قانون اجباری بستن کمربند در خودرو برداشته‌شود، به نظر شما چند نفر حاضرند تا کمربندشان ببیندند؟ (این خود نشان دهنده‌ی این است که مردم از سوگیری شناختی غفلت احتمالی استفاده می‌کنند!) این قضیه اصلا هم جالب نیست! زیرا «رویدادهایی با احتمال کم و تأثیر زیاد» اغلب دارای ارزش‌های مورد انتظار بزرگ، اعم از منفی یا مثبت هستند، این دام در تفکر انسان نگران‌کننده است! این نشان می‌دهد که انسان اغلب به «ارزش‌های مورد انتظار» حتی فکر هم نمی‌کند! چه برسد که بخواهد آن را هنگام تصمیم‌گیری به کار ببرد!

سوگیری دیگری که ممکن است بر توانایی افراد در برآورد مقادیر مورد انتظار تأثیر بگذارد، «غفلت از محدوده» است. مطالعات نشان داده است که افراد ارزش‌گذاری خود را در تناسب با مقیاس یک مسئله تنظیم نمی‌کنند. به عنوان مثال، یک مطالعه از سه گروه از افراد در مورد تمایل آن ها به پرداخت هزینه برای نجات 2000 یا 20000 یا 200000 پرنده از غرق شدن در استخرهای نفتی بدون سرپوش پرسیده شد. میانگین‌های مربوطه 80، 78 و 88 دلار و میانگین پاسخ‌ها همگی 25 دلار بود. اگر ارزش‌گذاری افراد از برخی نتایج به‌درستی مقیاس‌پذیر نباشد، ارزش‌های مورد انتظار

نیز نخواهد بود (یعنی اینجا باید هرکس با توجه به دارایی خود مبلغی را اعلام می‌کرد، ولی همه‌ی آن‌ها پاسخی نزدیک به 25 دلار داده‌بودند).

در انتخاب پروژه های خیریه، پروژه‌های (موارد) موثر را انتخاب کنید

از آنجایی که مردم معمولاً به جای تحقیق در مورد اثربخشی خیریه، بر اساس انگیزه و احساسات به خیریه می‌پردازند، اغلب از خیریه‌ها و اهداف بی‌اثر حمایت می‌کنند. ولی در عوض چیزاهایی نذیر: نوع دوستی مؤثر، یک جنبش جهانی اخیر، بر اهمیت رفتار نوع دوستانه مؤثر، چه در قالب کمک های مالی و چه در قالب زمان مهم هستند!

چه خوب است که همین اصل (سراغ کارهایی برویم که اثربخشی بالا دارند) را در هوش مصنوعی پیاده‌سازی کنیم و به اهداف مهم‌تر، اولویت بالاتری بدهیم.