

رفتار واقعی هوش مصنوعی برای دانشمندان داده

انتشارات محمد رحیمی

۱ ژوئن ۲۰۲۳

فهرست مطالب

| | |
|-----|---------------------------------------|
| ii | تشکر نامه |
| iii | فهرست همکاران و مشارکت کنندگان |
| ۱ | مقدمه: ماشین‌های اخلاقی |
| ۱ | ۱.۱ ماشین‌های اخلاقی |
| ۳ | ۲.۱ علم داده چیست؟ |
| ۴ | ۳.۱ موارد مطالعاتی |
| ۴ | ۱.۳.۱ مورد اول اخلاق تحقیق و روش علمی |
| ۵ | ۲.۳.۱ مدل‌های ماشین در دادگاه |
| ۷ | ۳.۳.۱ رسانه‌های ساختگی و خشونت سیاسی |
| ۹ | ۴.۳.۱ بیومتریک و تشخیصی چهره |
| ۱۰ | ۴.۱ بخش دوم |

تشکر نامه

مایلم از مشارکت کنندگان زیر برای کمک‌های عالی و متنوعشان در این کتاب تشکر کنیم: پیتر هرشوگ (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 متنوع و جذاب باشد. بدون کمک این عزیزان خردمند، توانستیم این کتاب را ایجاد کنیم!

فهرست همکاران و مشارکت کنندگان

همکاران:

جان مورونگی

گروه فلسفه دانشگاه وست چستر

پیتر سینگر

مرکز دانشگاهی برای ارزش های انسانی دانشگاه

پرینستون

مشارکت کنندگان:

جان هکر رایت

گروه فلسفه دانشگاه گوئلف

بیپ فای تسه

مرکز دانشگاهی برای ارزش های انسانی دانشگاه

پرینستون

دنیل سینکلر

دانشکده حقوق دانشگاه فوردھام

سامول جی لوین

مرکز حقوقی تورو

پیتر دی هرشوک

مرکز شرقی-غربی

کولین مارشال

گروه فلسفه دانشگاه واشنگتن

آندریا سالیوان کلارک

گروه فلسفه دانشگاه ویندزور

جویی میلر

گروه فلسفه دانشگاه وست چستر

فصل ۱

مقدمه: ماشین‌های اخلاقی

۱.۱ ماشین‌های اخلاقی

این کتاب، برای دانشمندان داده و افراد علاقه‌مند به این حوزه 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، کاملاً 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 درگیر 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 نوشته شده 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)، دین‌شناسی 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!

۲.۱ 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. مثلاً 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 زود هنگام 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 نسبتاً 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 قبلاً 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 کِلِهر (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!

۳.۱ موارد مطالعاتی

۱.۳.۱ 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 Lab Brand and Food Cornell به وسیله‌ی بریایان وانسینک Wansink Brian 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 (برای علنی کردن نتایجی که مثبت بودند)، 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.

۲.۳.۱ 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 تأیید شده نیز در زمینه‌های اجتماعی مختلف نیز عملکرد 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 قوی‌ای 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 جی 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 تصمیم‌گیری‌ها در سیستم عدالت کیفری ما نفوذ می‌کند. در همین حال، کسانی که به جای 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.

۳.۳.۱ 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 بعداً مشخص شد که ویدئو کاملاً واقعی بوده! در ویدئو به نظر می‌رسد که اثرات deepfake روی «بونگو» مشاهده می‌شود، ولی درواقع تأثیرات 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 می‌کنند، می‌پردازد. 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 می‌شوند. 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 نمی‌توان حقیقت را به‌طور 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 از دریچه اخلاق فایده‌گرایانه نگاه می‌کنند. آنها بر اهمیت حقیقت تأکید می‌کنند که (حقیقت) به معنای استفاده از روش علمی و شواهد تجربی معتبر برای تصمیم‌گیری 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 می‌دهند، با ارزش‌گذاری ما ساخته می‌شوند. پورنوگرافی چه به صورت کم عمق و چه به صورت عمیق، به‌طور ویژه سلامتی را تخریب می‌کند و توسط اخلاق «فایده‌گرایانه» 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. «آبه 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».^۱ 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 می‌کند 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 می‌دهند، تضعیف می‌کند.

«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.

۴.۳.۱ بیومتریک و تشخیص 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 اشکال 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 نقص قوانین حریم خصوصی شناخته‌شد.

۴.۱ بخش دوم

این یک مثال است.