مجله‌ی علمی پژوهش در علوم رایانه، شماره ۲۸، زمستان ۱۴۰۱، ص ۲۴- ۱۵

ISSN: 2745-3673
http://www.Rcsj.ir

مجله علمی پژوهش در علوم رایانه (سال هفتم)
شماره ۲۸ / زمستان ۱۴۰۱ / ص ۲۴- ۱۵

پژوهشی در خصوص بررسی اخلاق و مقررات در هوش مصنوعی

احسان باقری
دانشجوی کارشناسی ارشد مدیریت فناوری اطلاعات (موسسه آموزش عالی نورطوبی)

نام نویسنده مسئول:
احسان باقری

تاریخ دریافت: ۱۴۰۱/۰۹/۳۱
تاریخ پذیرش: ۱۴۰۱/۱۲/۰۲

چکیده

مطابق تعریف شرکت ای بی ام اخلاق در هوش مصنوعی مجموعه ای از دستورالعمل هایی است که در مورد طراحی و نتایج هوش مصنوعی توصیه می کند. این مقاله ابتدا سعی بر شناسایی و تاثیرات وجود اخلاق در هوش مصنوعی به عنوان یک عامل حیاتی برای یک کسب و کار دارد و پس از آن وضعیت موجود را از سه دیدگاه مقررات، اصطلاحات و انجمن‌ها و ابزارهای موجود بررسی میکند. نتایج حاصل از این موضوع حکایت دارد که با توجه به اینکه کم و بیش کارهای زیادی در این خصوص صورت گرفته است و این مباحث به شرکت‌های بزرگ محدود نیست اما هنوز تعاریف واحد منسجم و قراردادی مشخصی برای آن وجود ندارد و همچنین دخالت انسان در این موضوع در کلیه مراحل به عنوان یک راه حل مطرح شده است و نهایتا ابزارهایی هم برای نظارت بر این موضوع فهرست گردیده است که البته هیچ یک اجباری نیست.

واژگان کلیدی: اخلاق، هوش مصنوعی، یادگیری ماشین.

مجله‌ی علمی پژوهش در علوم رایانه، شماره ۲۸، زمستان ۱۴۰۱، ص ۲۴- ۱۵

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

هوش مصنوعی فرصت‌های بی‌سابقه‌ای را برای کسب‌وکارها به ارمغان می‌آورد، اما مسئولیت‌های باورنکردنی را نیز به همراه دارد. تأثیر مستقیم آن بر زندگی مردم سؤالات قابل توجهی را در مورد اخلاق هوش مصنوعی، حاکمیت داده، اعتماد و قانونمندی ایجاد کرده است. در واقع، تحقیقات نشان داد که از ۸۴ درصد از مدیران معتقدند که باید از هوش مصنوعی برای دستیابی به اهداف رشد خود استفاده کنند، با این حال ۷۶ درصد گزارش می دهند که با نحوه مقیاس بندی مشکل دارند و همچنین از هر چهار مدیر، سه نفر بر این باورند که اگر هوش مصنوعی را در پنج سال آینده مقیاس ندهند، خطر از بین رفتن کامل کسب و کار را خواهند داشت.[1] بیشتر کسب وکارها تلاش دارند، از سرمایه‌گذاری سهامداران محافظت نموده و تضمین بازگشت سرمایه مناسب، هم اکنون و هم در آینده وجود داشته باشد. این چیزی نیست به جز سود آوری یک شرکت. زمانی که سهامداران دید واضحی از آنچه در شرکت روی میدهد دارند و میتوانند اطمینان حاصل کنند که اقدامات لازم برای مشخص شدن و مدیریت هر مشکلی که ممکن است در آینده رخ دهد، وجود دارد، شرایط مطلوبی برقرار میباشد. اما سیستم‌های هوش مصنوعی اغلب غیر شفاف است و بیشتر فرایند تصمیم ناشناخته میماند. اینکه مردم واقعا به این مسئله آگاه باشند که آیا چیزی رو اتفاق میافتد چقدر منصفانه است میتواند مشخص کند که روی که چیزی پول خود را سرمایه گذاری کنند. از این جهت دولت ها به خوبی به این مسئله واقف هستند که هوش مصنوعی برای دولت‌ها در سرتاسر جهان اهمیت راهبردی پیدا کرده است و به عنوان یکی از دگرگون‌کننده‌ترین نیروهای زمان ما محسوب می‌شود. [2]. این پژوهش سعی دارد که تلاشهایی که برای ایجاد هوش مصنوعی منصفانه صورت گرفته را بررسی کند و نهایتا پیشنهاداتی در خصوص آن ارائه نماید.

با افزایش نگرانی های جهانی در مورد تأثیر هوش مصنوعی افسارگسیخته، تکنیک های توضیح پذیری اهمیت فزاینده ای پیدا می کنند. آنها راهی برای کاهش عدم اطمینان و کمک به جلوگیری از عواقب ناخواسته ارائه می دهند. تکنیک هایی که امروزه بیشتر مورد استفاده قرار می گیرند.

مرور ادبیات

عدم درک فنی از استفاده صحیح و مسئولانه از هوش مصنوعی از همان ابتدا توسعه را با مشکل مواجه می کند. با توجه به فقدان مهارت های علم داده در فضای فناوری اطلاعات به طور کلی و پیچیدگی برنامه نویسی هوش مصنوعی، این یک چالش رایج است[3] . حتی اگر سازمان‌ها علاقه متفاوتی به اصولی که باید روی آن تمرکز کنند نشان دهند، مرتبط‌ترین مفاهیمی که توسط سازمان‌ها و شرکت‌ها دیده می‌شود شامل حریم خصوصی، انصاف، پاسخگویی، شفافیت، توضیح‌پذیری، و ایمنی است[4] مون توضیح می‌دهد که مشکل مربوط به اختلاف نظر مفهوم در جوامع میباشد. ارزش‌های ستودنی مانند «انصاف» و «حریم خصوصی» زمانی که تحت بررسی قرار می‌گیرند از بین می‌روند و منجر به دیدگاه‌های متفاوت و اهداف عمیقاً ناسازگار می‌شوند [5] لویر تصریح میکند که چارچوب‌های اخلاقی هوش مصنوعی و دستورالعمل‌های پیاده‌سازی هوش مصنوعی باید کل محیط را که در آن این مؤلفه‌ها توسعه و مستقر شده‌اند (از جمله همه عوامل درگیر) را در نظر بگیرند [6].

دیگنوم و همکاران اصول اصل ART را پیشنهاد کردند[7] که شامل مسئولیت پذیری، مسئولیت و شفافیت یک رویکرد پایه برای اطمینان از اینکه ارزش ها و اصول اخلاقی در طراحی هوش مصنوعی گنجانده شده است میباشد.

انسان محوری را می توان یک زیر کلاس یا یک چارچوب هوش مصنوعی خاص در نظر گرفت که بر تعامل و همکاری با عوامل انسانی تمرکز دارد. تحت این چهارچوب، برای اینکه یک جزء هوش مصنوعی انسان محور در نظر گرفته شود، باید قابل توضیح و تفسیر باشد، قابل تأیید باشد (که می تواند به شش ویژگی عمومی مرتبط باشد: قابلیت اطمینان، ایمنی، در دسترس بودن، محرمانه بودن، یکپارچگی، و قابلیت نگهداری) [8]

آنهلکار و همکاران عنوان میکنند که درک عدم قطعیت‌های انسانی (از جمله سوگیری‌ها) توسط هوش مصنوعی به شما اجازه می‌دهد تا از دیدگاه هوش مصنوعی یک نمایش کامل از کل محیط سیستم داشته باشید. این می‌تواند با روش‌هایی امکان‌پذیر باشد که مسیر کاربر را پیش‌بینی می‌کنند[9]

ISSN: 2745-3673
http://www.Rcsj.ir

فریدمن و زلبرستین عنوان میکنند که درک عدم قطعیت‌های هوش مصنوعی توسط انسان‌ها که منجر به افزایش شیوه‌های ایمن با بهبود تفسیر نحوه عملکرد هوش مصنوعی می‌شود. راه‌حل‌های این مشکلات مستقیماً با ملاحظات قابل اعتماد بودن و شفافیت مرتبط است [10]

روش انجام کار

در این مقاله با استفاده از موتور جستجوی گوگل و همچنین استفاده از پایگاههای وب در sciencedirect و springer و با کلیدواژه های AI,ML, ethics بیش از یکصد مقاله بررسی شده و در زمینه تجربیات داخلی پایگاههای جهاد دانشگاهی و ایرانداک هم مورد جستجو قرارگرفت. معیار انتخواب مقالات با توجه به تاریخ آنها از آغاز ۲۰۱۸ و اولویت با مقالات انگلیسی میباشد.

یافته ها

چنانکه گفته شد در خصوص موضوع هوش مصنوعی منصف نظریات متفاوتی وجود دارد از اینکه انسان را ناظر فرض کنیم تا اینکه خطاها را یک آسب غیرعمدی بدانیم همه موضوعات مهم به شمار میروند اما پیچیدگی زمانی افزایش میابد که سازمان تصمیم به پیاده سازی یا قصد پیاده سازی مدلها را دارد. در این حالت آگاهی از محدودیتها قوانین و ابزارهای بهبود ضروری به نظر میرسد. به عنوان مثال احتمالا یک سازمان مالی علاقه ای به پیاده سازی الگوریتمهای شبکه عصبی را به دلیل کارکرد آنها نشان نمیدهد و تنها به درخت تصمیم اکتفا میکند تا دچار مواردی مثل تبعیض جنسیتی در اعطای وام نشود. در ادامه موارد و روشهایی مورد تاکید قرار گرفته است که بیشترین تاثیر را در این جریان دارند. به صورت کلی وضعیت را میتوان در سه بخش عمده بررسی کرد.

۱. قوانین
۲. اصطلاحات و بیانه ها و انجمنها
۳. ابزارها و پروژه ها

در ادامه این مقاله هر در خصوص جریانهای موجود در ارائه یک هوش مصنوعی اخلاقی به صورت جزیی بررسی شده است و نهایتا جمع بندی صورت گرفته است.

تعریف هوش مصنوعی و جریانهای سازنده آن

هوش مصنوعی در هسته خود، علم الگوریتم‌های کامپیوتری است که به‌جای برنامه‌ریزی صریح، به‌طور خودکار از تجربه یاد می‌گیرند و بهبود می‌یابند. الگوریتم‌ها داده‌های نمونه را که به داده‌های آموزشی معروف هستند، تجزیه و تحلیل می‌کنند تا یک مدل بسازند که بتواند پیش‌بینی کند. در واقع یک تابع قابل انطباق عرضه میشود که توسط ماشین میتواند آموزش ببیند. این مدلها میتوانند در صنایع مختلف استفاده شوند و نهایتا جهانی بدون کار، مرفه، عادلانه و معنادار ایجاد کنند. دنیایی که در آن همه به منابع و فرصت‌هایی برای پیشرفت دسترسی داشته باشند [11]. سوالی که در اینجا پیش میآیندبه این صورت است که موضوع داده ها به عنوان یکی از اجزای سازنده یکی از ایرادات به کل این تلاش این وضعیت است و موضوع مالکیت ممکن است به هیچ وجه راه درستی برای رفع نگرانی در مورد استفاده از داده نباشد [12] با این حساب مسئولیت مدل از آن چه کسی است؟

قوانین

امروزه مقررات کمی در سرتاسر جهان وجود دارد که به طور خاص با هدف یادگیری ماشین وضع شده باشند با این حال، بسیاری از مقررات موجود حاکمیت یادگیری ماشین تاثیر دارند.

- مقررات خاص صنعت. این امر به ویژه در بخش های مالی و داروسازی قابل توجه است.

ISSN: 2745-3673
http://www.Rcsj.ir

- مقررات طیف گسترده، به ویژه پرداختن به حریم خصوصی داده ها.

مقررات دارویی در ایالات متحده: GxP

GxP مجموعه‌ای از دستورالعمل‌های کیفیت و مقرراتی است که توسط سازمان غذا و داروی ایالات متحده (FDA) ایجاد شده است، که هدف آن اطمینان از ایمن بودن محصولات زیستی و دارویی است.

مقررات مدیریت ریسک مدل مالی

در امور مالی، ریسک مدل ریسک متحمل شدن زیان زمانی است که مدل‌های مورد استفاده برای تصمیم‌گیری در مورد دارایی‌های قابل معامله نادرست هستند. مقررات مدیریت ریسک مدل (MRM) با تجربه تأثیر رویدادهای خارق‌العاده، مانند سقوط‌های مالی، و آسیب‌های ناشی از آن به عموم و اقتصاد گسترده‌تر در صورت متحمل شدن زیان‌های شدید هدایت می‌شود. این نشان می دهد که چگونه موسسات مالی باید یک چارچوب مدیریت ریسک مدل جامع (MRM) را طراحی، اجرا و حفظ کنند.

مثلا SR11-7 یک استاندارد نظارتی است که توسط بانک فدرال رزرو ایالات متحده تنظیم شده است که راهنمایی هایی را در مورد مدیریت ریسک مدل ارائه می دهد. یا IRRBB CSRBB دستورالعمل‌های مربوط به ریسک نرخ بهره برای دفتر بانکی (IRRBB) و ریسک اعتبار ناشی از فعالیت‌های دفاتر غیرتجاری (CSRBB) جایگزین دستورالعمل‌های مربوط به جنبه‌های فنی مدیریت ریسک نرخ بهره ناشی از فعالیت‌های غیرتجاری تحت فرآیند بررسی نظارتی می‌شود.

مقررات حفظ حریم خصوصی داده های GDPR

مقررات عمومی حفاظت از داده های اتحادیه اروپا برای اولین بار در سال ۲۰۱۸ اجرا شد و دستورالعمل هایی را برای جمع آوری و پردازش اطلاعات شخصی افراد ساکن در اتحادیه اروپا تعیین کرد. با این حال، با در نظر گرفتن عصر اینترنت توسعه یافته است، بنابراین در واقع برای بازدیدکنندگان اتحادیه اروپا از هر وب سایت، صرف نظر از اینکه آن وب سایت در کجا قرار دارد، اعمال می شود. این مقررات در برخی از کشورهای دیگر مانند سنگاپور هم اجرایی شده است و یا در ایران پیشنهاد شده است [13].

جدول ۱- محدوده و مقررات

منطقه	مرحله
اتحادیه اروپا	راهنمایی، ارتباط، هدایت و مقررات
ایالات متحده	راهنمایی، ارتباطات و مقررات
انگلستان	راهنمایی
استرالیا	راهنمایی

اما کدام قوانین، در صورت وجود، ممکن است نقض شوند؟ اول از همه، واضح است که فناوری های دستکاری آزادی انتخاب را محدود می کنند. اگر کنترل از راه دور رفتار ما به خوبی کار می‌کرد، اساساً برده‌های دیجیتالی می‌شدیم، زیرا فقط تصمیماتی را که قبلاً توسط دیگران گرفته شده بود، اجرا می‌کردیم. البته فن آوری های دستکاری تنها تا حدی موثر هستند. با این وجود، آزادی ما به آرامی، اما مطمئناً در حال ناپدید شدن است – در واقع، آنقدر آهسته است که تاکنون مقاومت کمی از سوی جمعیت صورت گرفته است.[14]

ISSN: 2745-3673
http://www.Rcsj.ir

اصطلاحات و بیانه ها و انجمنها

با رشد علم داده و یادگیری ماشین هوش مصنوعی بین اندیشمندان این رشته اجماع و اعلاناتی برای استفاده از هوش مصنوعی صورت گرفته است. از جمله اینکه هوش مصنوعی باید پاسخگو، پایدار و قابل کنترل باشند. سیستم‌های هوش مصنوعی باید در طول زمان قابل اعتماد باقی بمانند و به خوبی کنترل شوند و همچنین قابل ممیزی باشند.

هوش مصنوعی مسئول

هیچ نهادی برای توافق و چهارچوب بندی هوش مصنوعی مسئول وجود ندارد و همینطور هیچ تعریف دقیقی برای این اصطلاحات وجود ندارد. در این زمینه، درک این نکته مهم است که هوش مصنوعی به خودی قادر به پایداری نیست، بلکه باید به عنوان بخشی از روابط اجتماعی و فنی درک شود. یک رویکرد مسئولانه به هوش مصنوعی مورد نیاز است [15]. به همین جهت در این موضوع شاهد پیدایش تعاریف مشخصی هستیم.

انسان در حلقه

مفهوم انسان در حلقه یک مفهوم کلی است که برای اعتبار سنجی مدلهای یادیگری ماشین در زمان آموزش استفاده میشود. مشخصا در مورد قضاوتهای اخلاقی انسانها بهتر از ماشین عمل میکنند و هدف اینجا درگیر کردن انسانها در مقاطع مختلف تولید مدل است.

سیستم‌های ترکیبی کاربرد محاسبات انسان در حلقه را برای کارآمدتر کردن فرآیندهای عملیاتی ارائه می‌کند. طراحان باید اعتماد را در رابطه انسان و هوش مصنوعی ایجاد و تقویت کنند تا برنامه‌های کاربردی آینده موفق و قابل اعتماد شوند. با روند رو به رشد پیشرفت‌های تکنولوژیکی و گنجاندن هوش مصنوعی در کاربردهای بیشتر و بیشتر، اتحاد انسان ها و ماشین ها حتی مهم تر شده است [16]

هدف از ادغام دانش حوزه انسانی نیز ارتقای خودکارسازی یادگیری ماشینی است. انسان در حلقه حوزه ای است که ما آن را در تحقیقات آینده اهمیت فزاینده ای می بینیم زیرا دانش آموخته شده توسط یادگیری ماشینی نمی تواند دانش حوزه انسانی را به دست آورد. انسان در حلقه با ادغام دانش و تجربه انسانی با هدف آموزش یک مدل پیش‌بینی دقیق با حداقل هزینه انجام می‌شود. انسان‌ها می‌توانند داده‌های آموزشی را برای برنامه‌های یادگیری ماشین فراهم کنند و به‌طور مستقیم وظایفی را که برای رایانه‌های در حال تولید سخت هستند، با کمک رویکردهای مبتنی بر ماشین انجام دهند [17]. ادغام فن‌آوری‌های هوش مصنوعی و سیستم‌های قابل یادگیری در تولید و تدارکات، مفاهیم سازمان‌دهی کار و انتساب وظایف را به عوامل انسانی و ماشینی تغییر می‌دهد. امروزه از این فناوری در بسیاری از سیستمها مانند تولید و تدارکات [18] ،پزشکی [19] [20] تجزیه تحلیل شبکه های اجتماعی و اخبار [21] استفاده میشود.

تکرار پذیری و ردیابی

تکرارپذیری در یادگیری ماشینی به این معنی است که می توان الگوریتم را مکررا بر روی مجموعه داده خاصی اجرا کرد و نتایج مشابه را در یک پروژه خاص به دست آورد. عنوان مثال تمرکز و ردیابی کارها بر گسترش درک سهامداران بخش تولید از رویکردهایی است که می‌توان در چرخه عمر هوش مصنوعی در نظر گرفت، شکاف بین اصول و الزامات قابل اجرا را کاهش داد و ملاحظات اساسی مبتنی بر مدیریت ریسک را برای ترکیب و مدیریت هوش مصنوعی مبتنی بر مسئولیت‌پذیر تعریف کرد [22]. به این منظور ذخیره اطلاعات در مورد شرایط و محیط توسعه مدل برای اینکه بتوان با همان نتایج از ابتدا مدل را بازتولید نمود از الزامات به شمار می آید. یکی از آشکارترین پیامدها تکرارپذیری ناکافی است که در مدل‌های پیش‌بینی کنونی یافت می‌شود که موجب بروز خطاهای زیادی میشود [23] ایجاد تکرار مستلزم ایجاد مجدد تمام شرایط مانند داده ها و پارامترها میباشد.

ISSN: 2745-3673
http://www.Rcsj.ir

انسان خارج از حلقه

اولین مرجع منتشر شده در مورد انسان خارج از حلقه در سال ۱۹۶۳ شی مدیر برنامه آپولو است. او به «مرد خارج از حلقه» به عنوان یکی از چندین اصطلاح مشهور که قبلاً استفاده می‌شد به کار میبرد. او این موضوع را در مورد اینکه یک مرد می تواند در چه وظایفی به بهترین شکل انجام دهد، بحث می کند. سندرز این موضوع را به عنوان یک اصطلاح نظامی برای حذف کامل انسان از یک سیستم کنترل نظامی توصیف میکند [24] پس از آن توماس و پریسن به مشکل از دست دادن آگاهی موقعیتی انسانها که کنترل را از یک موقعیت کاملاً خودکار «خارج از حلقه» بازپس می گیرند اشاره میکنند که نشان می دهد انسان خارج از حلقه مانع از نظارت انسان نمی شود [25].

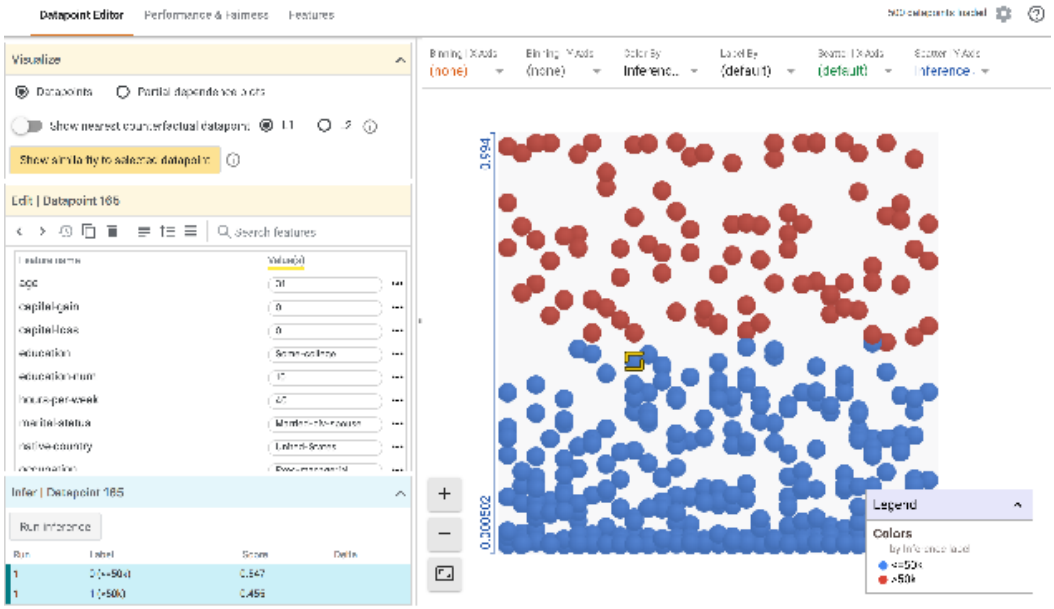
در سال ۲۰۱۸ مارات و همکاران تلاش می کنند تا انسان خارج از حلقه را در زمینه وسایل نقلیه خودکار به این صورت تعریف کنند: "نه در کنترل فیزیکی وسیله نقلیه، و نه نظارت بر وضعیت رانندگی، یا در کنترل فیزیکی وسیله نقلیه اما نه نظارت بر وضعیت رانندگی [26]

ابزارها و پروژه ها

ابزارهای اخلاقی می توانند به برنامه ها کمک کنند تا سیستم های مبتنی بر داده ها منصفانه تر، قوی تر و شفاف ترتولید شوند در ادامه چند مهم فهرست شده است.

ابزار WIF

از جمله شرکتهای پیش رو در زمینه مقررات هوش مصنوعی منصف گوگل است. گوگل ابزاری به نام What-If طراحی کرده است. هدف آن این است که به کاربر اجازه دهد یک مدل ML را با حداقل کدگذاری مورد نیاز بررسی کنید. با این ابزار می توانید داده ها و خروجی مدل را برای آن داده ها با هم بررسی کرد. ابزار What-If به کاربر امکان می دهد سیگنال های مختلفی مانند این را آزمایش کند، از جمله جزئیاتی مانند جنسیت، نژاد و موارد دیگر. این ابزار شامل توانایی پیدا کردن نزدیکترین موارد خلاف واقع میباشد و نزدیک‌ترین مجموعه داده را که منجر به استنتاج متفاوتی می‌شود پیدا میکند.



شکل- ابزار WIT

ابزار What-If (WIT) یک رابط کاربری آسان برای گسترش درک طبقه بندی جعبه سیاه و مدل های ML ارائه می دهد. با استفاده ازآن، می توان استنتاج را روی مجموعه بزرگی از داده انجام داد و بلافاصله نتایج را به روش های مختلف نمایش داد.

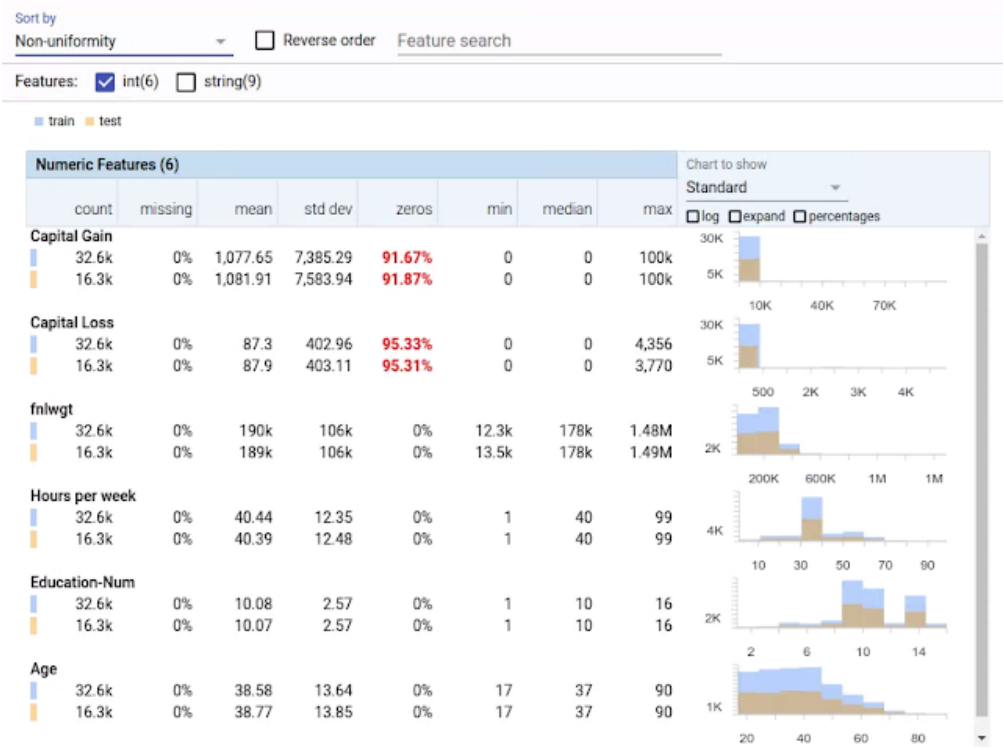
ISSN: 2745-3673
http://www.Rcsj.ir

علاوه بر این، دادها را می توان به صورت دستی یا برنامه نویسی ویرایش کرد و مجدداً از طریق مدل اجرانمود این شامل ابزاری برای بررسی عملکرد مدل و انصاف در زیر مجموعه های یک مجموعه داده است.

ابزار Facets

گوگل ابزار دیگری هم به نام Facets معرفی کرده است. Facets ابزاری است که می‌تواند مکمل What-If Tool باشد تا از طریق تجسم‌سازی به کاربر اطلاعات عمیقی در مورد داده‌ها بدهد. هدف Facets کمک به درک توزیع مقادیر در بین ویژگی‌های مجموعه داده میباشد.

Facets از دو بخش تصویرسازی تشکیل شده است که به کاربران اجازه می دهد تصویری کلی از داده های خود را در جزئیات مختلف ببینند. با استفاده ازFacets Overview، شکل هر ویژگی داده را درک کنید یا با استفاده از Facets Dive مجموعه‌ای از مشاهدات فردی را کاوش کنید. این تجسم‌ها امکان می‌دهند داده‌ها اشکال زدایی شوندfacets Dive. یک رابط کاربری و بصری برای کاوش در رابطه بین نقاط داده در میان ویژگی های مختلف یک مجموعه داده را فراهم میکند.



شکل ۲- ابزار Facets

پروژه Fairlearn توسط مایکروسافت ارائه شده است و بر این اساس استوار است که ناعادلانه بودن در سیستم‌ها را به حداقل برساند با توجه به اینکه تمایز بین دلایل ناعادلانه بودن می تواند دشوار باشد، به خصوص که آنها متقابلاً منحصر به فرد نیستند و اغلب یکدیگر را تشدید می کنند. بنابراین، این وظیفه به عهده انسان گذاشته میشود که آیا یک سیستم هوش مصنوعی از نظر تأثیر آن بر افراد و نه از نظر دلایل خاص مانند سوگیری‌های اجتماعی یا از نظر قصد، مانند تعصب، غیرمنصفانه رفتار می‌کند یا خیر

Fairlearn یک پروژه منبع باز و جامعه محور است که به دانشمندان داده کمک می کند تا عدالت سیستم های هوش مصنوعی را بهبود بخشند. توسعه Fairlearn کاملاً مبتنی بر درک این است که عدالت در سیستم‌های هوش مصنوعی یک چالش اجتماعی و فنی است. از آنجایی که منابع پیچیده بی‌عدالتی - برخی اجتماعی و برخی فنی - وجود دارد، نمی‌توان به طور کامل یک سیستم را "نقض" کرد یا انصاف را تضمین کرد. این پروژه از یک کتابخانه پایتون برای ارزیابی و بهبود عادلانه

ISSN: 2745-3673
http://www.Rcsj.ir

(متریک های انصاف، الگوریتم های کاهش، ترسیم و غیره) و منابع آموزشی که فرآیندهای سازمانی و فنی را برای کاهش بی عدالتی پوشش می دهد) راهنمای جامع کاربر، مطالعات موردی دقیق و سایر مواردتشکیل شده است.

پروژه متن بار Deon

این پروژه یک لیست خلاقی برای دانشمند علوم داده پاسخگو ایجاد میکند. Deon نقطه شروعی برای تیم ها برای ارزیابی ملاحظات مربوط به برنامه های کاربردی تجزیه و تحلیل پیشرفته و یادگیری ماشین از جمع آوری داده ها از طریق استقرار است. این پروژه با زبان پایتون است.

پروژه Model Cards

این پروژه توسط بخش تحقیقات گوگل معرفی شده است و به این سوال میپردازد که مدل تحت چه شرایطی بهترین و پایدارترین عملکرد را دارد؟ آیا نقاط کور دارد؟ اگر چنین است، کجا؟ به طور سنتی، پاسخ به چنین سؤالاتی به طرز شگفت آوری دشوار است. در حال حاضر، کارت‌های مدل گوگل مستنداتی در مورد عملکرد و محدودیت‌های یک الگوریتم ارائه می‌کنند. سایر موارد، مانند عملکرد مدل در ابعادی مانند نژاد و جنسیت نیز باید گنجانده شود تا سوگیری ضمنی آشکار شود.

پروژه AI Fairness 360

این پروژه توسط ای بی ام توسعه داده شده است و شامل جعبه ابزار متن باز قابل توسعه است که می تواند در بررسی، گزارش، و کاهش تبعیض و تعصب در مدل های یادگیری ماشین در طول چرخه عمر برنامه هوش مصنوعی کمک کند. بسته AIF360 نرم افزاری شامل معیارهای مختلفی است .علاوه بر این، ایی بی ام الگوریتم هایی را برای کاهش تعصب در مجموعه داده ها و مدل ها ارائه می دهد. کاربرد جعبه ابزار چند وجهی و وابسته به زمینه است.

نتیجه گیری

حتی اگر داده عالی باشند، یک سیستم مهندسی ضعیف می تواند منجر به مشکلاتی شود. عجله کردن در ارائه به بازار با یک محصول زمانی که لزوماً نیازی به آن نیست، یا زمانی که اطلاعات کافی برای ساختن یک محصول وجود ندارد. می‌تواند شرکت را در مسیر ساخت مدل‌های مغرضانه و متحمل شدن بدهی‌های فنی سنگین در آینده سوق دهد.به رغم عدم وجود قوانین نهاد نظارتی مشخص و حتی تعاریف یکسان اتحادیه اروپا و شرکتهایی مانند گوگل و میکروسافت در وضع قوانین و اصول هوش مصنوعی منصف پیشرو میباشند و غالبا برای این موضوع معیارهایی چون سودمندی اجتماعی، مبارزه با تعصب ناعادلانه، آزمایش کافی قبل از استفاده و پاسخگویی با حفظ حریم خصوصی مورد توجه قرار میگیرد. به هر حال باید توجه داشت که خطوط راهنما توسط کمپانیهای بزرگ در کنار ابزار درک مفهوم مدل را ساده تر مینماید ولی تمامی ابهامات را از بین نمیبرد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] Athena Reilly, Joe Depa and Greg Douglass, "AI: BUILT TO SCALE," Accenture, 2019.
[2] "COM(2018)237 - Communication," in Communication Artificial Intelligence for Europe, 2018, p. 1.
[3] T. Brown, "The AI skills shortage," IT chronicles, 2020. [Online]. Available: https://itchronicles.com/artificial-intelligence/the-ai-skills-shortage./
[4] R. Eitel-Porter, "Beyond the promise: implementing ethical AI," AI and Ethics, p. pages 73–80, 2021.
[5] L. Munn, "The uselessness of AI ethics," AI and Ethics, 2022 .
[6] D. Lauer, "You cannot have AI ethics without ethics," AI and Ethics, pp. 21-25, 2021 .
[7] Frank Dignum, Virginia Dignum, Javier Vázquez-Salceda, Aurélie Clodic, Manuel Gentile, Samuel Mascarenhas, and Agnese Augello, "Design for Values for Social Robot Architectures," HAL open science, 2018.
[8] J. Knight, Fundamentals of Dependable Computing for Software Engineers, CRC Press, 2012.
[9] Vaibhav V. Unhelkar, Claudia P´erez-D'Arpino, Leia Stirling and Julie A. Shah, "Human-robot co-navigation using anticipatory indicators of human walking motion," IEEE, 2015.
[10] R. Freedman and S. Zilberstein, "Safety in AI-HRI: Challenges Complementing User Experience Quality," AAAI Fall Symposia, 2016.
[11] Daniel Susskind, A World Without Work: Technology, Automation, and How We Should Respond, Metropolitan Books, 2020.
[12] M. Risse, "Data as Collectively Generated Patterns:," Carr Center for Human Rights Policy Harvard Kennedy School, Harvard University, 2021.
[13] Hossein Sadeghi and Mahdi Naser, "Legal-Ethical Challenges of the EU regulation on Artificial Intelligence," Bioethics Journal, vol. 10, no. 35, 2020.
[14] B. S. F. G. G. E. H. M. H. Y. H. J. v. d. H. R. V. Z. A. Z. Dirk Helbing, "Will Democracy Survive Big Data and Artificial Intelligence?," scientificamerican, 2017.
[15] Virginia Dignum, Responsible Artificial Intelligence How to Develop and Use AI in a Responsible Way, Springer Nature Switzerland, 2019.
[16] Julia Ostheimer, Soumitra Chowdhury and Sarfraz Iqbal, "An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles," Technology in Society, vol. 66, 2021.
[17] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma and Liang He, "A survey of human-in-the-loop for machine learning," Future Generation Computer Systems, vol. 135, pp. 364-381, 2022.
[18] Steffen Nixdorf, Minqi Zhang, Fazel Ansari and Eric H. Grosse, "Reciprocal Learning in Production and Logistics," IFAC-PapersOnLine, vol. 55, no. 2, pp. 854-859, 2022.
[19] Akash Gupta, Michael T. Lash and Senthil K. Nachimuthu, "Optimal sepsis patient treatment using human-in-the-loop artificial intelligence," Expert Systems with Applications, vol. 169, 2021.
[20] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes and Jacinto C. Nascimento, "BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions," Artificial Intelligence in Medicine, vol. 127, 2022.
[21] Rahul Pandey, Hemant Purohit, Carlos Castillo and Valerie L. Shalin, "Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning," International Journal of Human-Computer Studies, vol. 160, 2022 .
[22] Eduardo Vyhmeister, Gabriel G. Castane, Johan Buchholz and Per-Olov Östberg, "Lessons learn on responsible AI implementation: the ASSISTANT use case," IFAC-PapersOnLine, vol. 55, no. 10, pp. 377-382, 2022 .
[23] T. D. Azad, J. Ehresman, A. K. Ahmed, V. E. Staartjes, D. Lubelski, M. N. Stienen, A. Veeravagu and J. K.Ratliff, "Fostering reproducibility and generalizability in machine learning for clinical prediction modeling in spine surgery," The Spine Journal, vol. 21, no. 10, pp. 1610-1616, 2021 .
[24] D. S. Sanders, "Social Work Concerns Related to Peace and People Oriented Development in the International Context," Journal of Sociology and Social Welfare, 1988 .

ISSN: 2745-3673
http://www.Rcsj.ir

[25] T. Porathe and Johannes Prison, "Design of human-map system interaction," Extended Abstracts on Human Factors in Computing Systems, 2008 .

[26] B. S. T. L. J. E. J. D. L. E. J. C. A. G. S. K. C. M. M. I. D. M. T. S. K. U. T. V. A. S. &. A. K. Natasha Merat, "The “Out-of-the-Loop” concept in automated driving: proposed definition, measures and implications," Cognition, Technology & Work, vol. 21, p. 87–98, 2019 .

[27] N. S. N. do Faran and Eli (Omid) David, "Ground Truth Simulation for Deep Learning Classification of Mid-Resolution Venus Images Via Unmixing of High-Resolution Hyperspectral Fenix Data," IEEE International Geoscience and Remote Sensing Symposium, 2019.