كيف يمكن التغلب على تحيز الذكاء الاصطناعي... تقنييات وأدوات

مكّنت أنظمة الذكاء الاصطناعي الأشخاص في جميع أنحاء العالم من اختبار تجارب وقدرات جديدة، فاستُخدمت على نطاق واسع في مجالات عدّة، بداية من ترشيحات الكتب والبرامج التليفزيونية وحتى القيام بمهام أكثر تعقيدًا كاختيار المُرشّحين للالتحاق بالوظائف المختلفة، أو حتى مهام أكثر حساسية كالتنبؤ بالإصابة بالأمراض.

تحتاج أنظمة الذكاء الاصطناعي إلى كميات كبيرة من البيانات لتزداد دقتها في تنفيذ المهمات المطلوبة، وبعد الجمع والتخزين يتم معالجة تلك البيانات بواسطة خوارزميات النظام الذكية حتى يتعلم النظام من أنماط وخصائص البيانات، ومن ثم تطوير قدرة الآلة على إنجاز المهام التي صُنعِت من أجلها.

تحيز الذكاء الاصطناعي يمكن تعريفه على أنه حالة انحراف في نتائج خوارزميات التعلم الآلي يحدث بسبب وجود فرضيات مُتحيّزة أثناء عملية تطوير الخوارزمية، وتمثل انعكاسًا لعنصرية المجتمع وتحيزه ضد فئة معينة، أو قد يكون نتيجة تحيز في بيانات التدريب التي يتم تغذية نظام الذكاء الاصطناعي بها.

مع انتشار تطبيقات الذكاء الاصطناعي في القطاعات والمجتمعات المختلفة ازداد تأثير تحيزها بشكل واسع النطاق، ومعه ازدادت الحاجة إلى أنظمة عادلة قادرة على اتخاذ قرارات – أو المساعدة في اتخاذها- تتسم بالإنصاف ومُجرّدة من أشكال العنصرية والتحيز.

ممارسات للحد من تحيز الذكاء الاصطناعي

تطوير أنظمة ذكاء اصطناعي تتسم بالإنصاف وتنأى عن التحيز ليست مهمة سهلة لعدة أسباب؛ ربما تكون أهمها هو تغذية نماذج التعلم الآلي بالبيانات التي تُجمع من العالم الحقيقي، وبالتالي فإن أكثر الأنظمة دقة يمكن أن تتعلم أو تضخم التحيزات الموجودة مسبقًا في هذه البيانات، والتي يمكن أن تحتوي على تحيّز قائم على أساس العرق أو الجنس أو الدين أو أي خصائص أخرى. يمكن أيضًا أن يكشف النظام عن نقاط عمياء غير مقصودة بعد إطلاقه، هذه النقاط قد تكون إشكاليات تظهر قبل أو أثناء أو بعد تطوير نظام الذكاء الاصطناعي نتيجة وجود تحيزات أو أحكام مسبقة أو تفاوتات هيكلية في المجتمع، وتحدث حتى مع تدريبه واختباره بشكل صارم. على سبيل المثال، إذا تم تدريب نظام الذكاء الاصطناعي للتعرف على أصوات البالغين فإنه سيكون عادلًا وشاملًا في هذا النطاق، ولكنه قد يفشل في التعرف على الكلمات أو العبارات العامية المستحدثة إذا تم استخدامه من قِبل المراهقين.

لا يوجد تعريف موحد للإنصاف، سواء كان اتخاذ القرار يتم بواسطة البشر أو الآلات، فتحديد معايير الإنصاف المناسبة لنظام ما يتطلب مراعاة تجربة المستخدم والاعتبارات الثقافية والاجتماعية والتاريخية والسياسية والقانونية والأخلاقية، والتي قد يكون للعديد تفضيلات مختلفة اتجاهها.

في محاولة للتغلب على مشكلة تحيز الذكاء الاصطناعي نشرت شركة Innodata المتخصصة في هندسة البيانات 5 ممارسات يمكن الاعتماد عليها للحد من تحيز الذكاء الاصطناعي والوصول إلى نماذج تعلم آلي أكثر إنصافًا وشمولًا، والممارسات هي:

- اختيار مجموعة البيانات

في حين أن التخفيف من حدة تحيز الذكاء الاصطناعي والتعلم الآلي يمكن أن يمثل تحديًا، إلا أن هناك تقنيات وقائية يمكن أن تساعد في حل هذه المشكلة. التحدي الأكبر في تحديد التحيز يتمثل في فهم كيف تقوم بعض خوارزميات التعلم الآلي بتعميم البيانات التي تم تدريبها عليها. لذلك من المهم أن تتسم البيانات المستخدمة لتدريب النموذج بالشمول.

- تنوع الفريق

بناء فريق متنوع يساهم إلى حد كبير في القضاء على التحيز. يمكن أن يكون لتنوع الفريق تأثير إيجابي على نماذج التعلم الآلي من خلال إنتاج مجموعات بيانات مُمثّلة ومتوازنة (Representative Datasets). كما أنه يساعد في التخفيف من التحيز الضار في بنية مجموعات البيانات وكيفية تطبيق التصنيفات على تلك البيانات.

- تقليل الإقصاء

يعد اختيار الميزة (Feature selection) أمرًا أساسيًا للمساعدة في تقليل الإقصاء في الذكاء الاصطناعي، وهي عملية تقوم على تقليل عدد المتغيرات المُدخلّة إلى نماذج الذكاء الاصطناعي بهدف تحسين أداء التنبؤ بالنتائج. تستبعد هذه الخطوة عناصر البيانات التي لا تحتوي على تباين كافٍ للتأثير على النتائج.

- الخوارزميات وحدها ليست كافية

هناك طريقة أخرى للبدء في حل مشكلة التحيز في الذكاء الاصطناعي، وهي عدم الاعتماد فقط على الخوارزميات، بل إبقاء الأفراد القائمين على تطوير الذكاء الاصطناعي على اطلاع بكل ما يتعلق بالنظام ليتمكنوا من التعرف بفعالية على أنماط التحيز غير المقصود. يمكن أن تعمل هذه الخطوة على تقليل العيوب الموجودة في النظام، مما يخلق نموذج تعلم آلي أكثر حيادية. يجب على المنظمات أيضًا وضع مبادئ توجيهية وإجراءات تحدد التحيز المحتمل في مجموعات البيانات وتخفف من حدته. يمكن أن يساعد توثيق حالات التحيز عند حدوثها وتحديد كيفية العثور عليها والحديث بشأنها، على ضمان عدم تكرارها.

- بيانات ممثلة

يجب على المؤسسات فهم كيف تبدو البيانات الممثلة قبل جمع البيانات التي سيتم تدريب نموذج التعلم الآلي عليها. كما يجب أن يحتوي جوهر وخصائص البيانات المستخدمة أقل قدر من التحيز. وإلى جانب تحديد التحيز المحتمل في مجموعات البيانات، يجب على

المؤسسات أيضًا توثيق أساليبها في اختيار البيانات وتنقيتها، للقضاء جذريًا على أسباب التحيز.

اختبار التذوق الأعمى والذكاء الاصطناعي

بالإضافة إلى الممارسات السابقة للحد من تحيز الذكاء الاصطناعي، يمكن أيضًا الاعتماد على ما يعرف باختبار التذوق الأعمى، وهو اختبار يتم استخدامه منذ عقود. انتشر هذا الاختبار في منتصف سبعينات القرن الماضي، عندما أطلقت إحدى شركات المشروبات الغازية تحدي يعتمد على تذوق نوعين من المشروبات الغازية دون معرفة الشركة المُنتجة للمشروب، عن طريق إزالة الملصق عن عبوات المشروب. جاءت نتيجة التحدي لصالح الشركة صاحبة التحدي الذي فضلته الأغلبية على المنافس الأكثر مبيعًا، وذلك على الرغم من أن ملصق المنافس كان يخلق تحيزًا لصالح المنتج على أرض الواقع.

أثبتت التجربة السابقة أن إزالة معلومات التعريف (الملصق) عن المنتج أزالت معها التحيز وبالتالي اعتمد الناس في اختيارهم على التذوق فقط، وهو ما يمكن تطبيقه أيضًا على الآلات باستخدام اختبار التذوق الأعمى. يعني هذا أنه يمكن للخوارزميات ببساطة أن ترفض المعلومات التي قد تتسبب في إنتاج مخرجات متحيزة، وذلك للتأكد من أن نموذج التعلم الآلي يصنع تنبؤات "عمياء" عن هذه المعلومات.

كمثال على ذلك، تستخدم مقاطعة بنسلفانيا أداة تدعى Allegheny Family Screening Tool (AFST) للتنبؤ باحتمالية تعرض الأطفال لمواقف مسيئة. تعتمد الأداة على الذكاء الاصطناعي عبر استخدام بيانات من إدارة الخدمات الإنسانية بالمقاطعة، وتشمل تلك البيانات سجلات الهيئات العامة المتعلقة برفاهية الأطفال والخدمات الخاصة بإساءة استخدام الكحول والمخدرات والإسكان وغيرها. يستخدم العاملون في مجال دراسة الحالات بلاغات عن تعرض أطفال لإساءة محتملة، إلى جانب أي بيانات متاحة للعامة وتخص العائلة المشتبه بها، وذلك لتشغيل نموذج التعلم الآلي الذي يتنبأ بمستوى المخاطر من 1 إلى 20، ويتم فتح تحقيق إذا كانت الدرجة التي تنبأ بها النظام مرتفعة.

ولكن أدمج نظام AFST تحيزات بشرية في نموذج الذكاء الاصطناعي. أحد أكبر هذه التحيزات هو أن النظام يضع في الاعتبار المكالمات المتعلقة بالعائلات من مقدمي الرعاية الصحية إلى الخط الساخن لخدمة المجتمع. وتشير بعض الأدلة على أن مثل هذه المكالمات أكثر احتمالًا بثلاث مرات أن تتعلق بالعائلات ذات البشرة السمراء أو ثنائية العرق مقارنة بالعائلات ذات البشرة البيضاء. وعلى الرغم من استبعاد العديد من هذه المكالمات في نهاية المطاف، إلا أن النظام يعتمد عليها في تحديد درجة المخاطر، مما يؤدي إلى فتح تحقيقات متحيزة عنصريًا إذا كان المتصلون على الخط الساخن أكثر احتمالية للإبلاغ عن العائلات ذات البشرة السمراء.

في هذه الحالة، يمكن أن يعمل "اختبار التذوق الأعمى" على النحو التالي: تدريب النموذج على جميع البيانات التي يمكن استخدامها في التنبؤ باحتمالية تعرّض أطفال للإساءة، بما في ذلك المكالمات المحالة من الخط الساخن لخدمة المجتمع. ثم إعادة تدريب النموذج على جميع البيانات باستثناء هذا العامل. إذا كانت تنبؤات النموذج جيدة بنفس القدر بغض النظر عن عامل المكالمات، فهذا يعني أن النموذج يقوم بتنبؤات عمياء. أما إذا كانت التنبؤات مختلفة عند تضمين هذه المكالمات، فهذا يشير إلى أن المكالمات تمثل متغيرًا توضيحيًا في النموذج، أو قد يكون هناك تحيز محتمل في البيانات التي يجب فحصها بشكل أكبر قبل الاعتماد على الخوارزمية.

أدوات للتغلب على تحيز الذكاء الاصطناعي

- IBM AI Fairness 360: قدمت IBM Research مجموعة أدوات "AI Fairness 360 وAIF360" في 2018، وهي مجموعة من المقاييس الشاملة مفتوحة المصدر للتحقق من التحيز غير المرغوب فيه في قواعد البيانات ونماذج التعلم الآلي. تتشكل هذه الأدوات من مجموعة من الخوارزميات الحديثة التي يمكنها التخفيف من تحيز الذكاء الاصطناعي. احتوى الإصدار الأولي من حزمة AIF360، المتاحة كحزمة Python، على تسعة خوارزميات مختلفة للتخفيف من التحيز غير المرغوب فيه. حزمة AIF360 ليست فقط مجموعة من الأدوات، لكنها تحتوى أيضًا على تجربة تفاعلية توفر مقدمة بسيطة عن مفاهيم وإمكانيات الحزمة، وذلك للمساعدة على معرفة المقاييس والخوارزميات الأكثر ملائمة لحالة معينة. كما

إنها صُممت لتكون مفتوحة المصدر لتشجيع مساهمة الباحثين من جميع أنحاء العالم لإضافة المقاييس والخوارزميات الخاصة بهم إلى الحزمة. اتسم الفريق الذي عمل على إصدار الحزمة بالتنوع من حيث المجموعة العرقية، والاختصاص العلمي، والهوية الجنسية، وسنوات الخبرة، ومجموعة من الخصائص أخرى.

- Fairlearn: هي مجموعة أدوات مفتوحة المصدر، مُقدّمة من Microsoft، تُمكِّن علماء البيانات والمطورين من تقييم وتحسين الإنصاف في أنظمة الذكاء الاصطناعي الخاصة بهم. يحتوي Fairlearn على مكونين: لوحة تحكم تفاعلية وخوارزميات للحد من التحيز. كما يطمح المشروع ليشمل مكتبة Python لتقييم مدى الإنصاف في الذكاء الاصطناعي وتحسينه (مقاييس الإنصاف، وخوارزميات التخفيف من التحيز، وما إلى ذلك)، وكذا الموارد التعليمية التي تغطي العمليات التنظيمية والتقنية للحد من التحيز في الذكاء الاصطناعي (دليل مستخدم شامل، دراسات حالة مفصلة، تقارير فنية، إلخ). ارتكز تطوير حزمة أدوات Fairlearn على حقيقة أن تحقيق الإنصاف في أنظمة الذكاء الاصطناعي يمثل تحديًا مجتمعيًا تقنيًا، نظرًا لوجود العديد من مصادر التحيز المعقدة التي يكون بعضها مجتمعي وبعضها تقني. لذلك، تم إنشاء وتطوير حزمة Fairlearn مفتوحة المصدر للسماح للمجتمع بالكامل -بدءًا من علماء البيانات والمطورين وصناع القرار في مجال الأعمال ووصولًا إلى الأشخاص الذين قد تتأثر حياتهم بتنبؤات أنظمة الذكاء الاصطناعي- بالمشاركة وتقييم الأضرار المتعلقة بالتحيز، ومراجعة تأثيرات استراتيجيات الحد من التحيز، ومن ثم جعلها مناسبة لسيناريوهاتهم.
- FairLens: هي أحد مكتبات Python مفتوحة المصدر والتي تُستخدم لاكتشاف التحيز وقياس الإنصاف في البيانات بشكل تلقائي. يمكن لحزمة Fairlens تحديد التحيز بشكل سريع، كما توفر مقاييس متعددة لقياس الإنصاف عبر مجموعة من الخصائص المحمية قانونًا مثل العمر والعرق والجنس. يمكن تلخيص السمات الأساسية لأداة Fairlens في أربعة نقاط، هي:
  - إمكانية قياس مدى التحيز: تحتوى الأداة على مقاييس واختبارات يمكن من خلالها تحديد مدى ودلالة التحيز باستخدام قياسات ومسافات إحصائية.
  - إمكانية اكتشاف الخصائص المحمية: توفر الأداة طرق لاكتشاف الصفات المحمية قانونًا، وتُمكِّن المستخدم من قياس العلاقات الخفية بين هذه الصفات وغيرها.

- أدوات التمثيل المرئي للبيانات: توفر Fairlens رسوم بيانية عن الأنواع المختلفة من المتغيرات في مجموعات فرعية من البيانات الحساسة، لتوفر بذلك طريقة سهلة لرؤية وفهم الاتجاهات والأنماط الموجودة في البيانات.
- تقييم الإنصاف: وهي طريقة مبسطة لتقييم عدالة مجموعة بيانات عشوائية، وإنشاء تقارير تسلط الضوء على التحيزات والعلاقات الخفية.

- Aequitas: هي مجموعة أدوات مرنة مفتوحة المصدر تعمل على التدقيق في تحيز الذكاء الاصطناعي، طور مجموعة الأدوات مركز علوم البيانات والسياسة العامة في جامعة شيكاغو بالولايات المتحدة. يمكن استخدام مجموعة الأدوات لمراجعة تنبؤات أدوات تقييم المخاطر المستخدمة في أنظمة العدالة الجنائية، والتعليم، والصحة العامة، وتنمية القوى العاملة والخدمات الاجتماعية القائمة على التعلم الآلي لفهم الأنواع المختلفة من التحيزات، واتخاذ قرارات مستنيرة بشأن تطوير ونشر مثل هذه الأنظمة، كما تساعد مجموعة الأدوات على تحديد أماكن وجود التحيزات في نموذج التعلم الآلي. يمكن من خلال مجموعة الأدوات الكشف عن نوعين من التحيزات في نظام تقييم المخاطر، وهما:
  - الإجراءات أو التدخلات المتحيزة التي لم يتم تخصيصها بشكل يمثّل كافة السكان.
  - المخرجات المتحيزة الناتجة عن خطأ في النظام تجاه مجموعات معينة من الأشخاص.
- TCAV: هو نظام أعلن عنه سوندار بيتشاي، الرئيس التنفيذي لشركة Google خلال مؤتمر Google I/O لعام 2019، وهو أيضًا مبادرة بحثية حملت اسم (Test With Concept Activation (TCAV للكشف عن التحيز في نماذج التعلم الآلي. يمكن للنظام فحص النماذج لتحديد العناصر التي يمكن أن تؤدي إلى التحيز على أساس العرق والدخل والموقع وما إلى ذلك. يتعلم نظام TCAV "المفاهيم" بشكل أساسي عن طريق الأمثلة.
- Google What-If Tool: أنشأ الباحثون والمصممون في Google أداة What-If، والتي تعني "ماذا لو"، كمورد عملي لمطوري أنظمة التعلم الآلي. تحاول الأداة الإجابة على أحد أكثر الأسئلة صعوبة وتعقيدًا فيما يتعلق بأنظمة الذكاء الاصطناعي، وهو "ما هو الإنصاف الذي يريده المستخدمون؟". تسمح هذه الأداة التفاعلية مفتوحة المصدر للمستخدم بالتحقيق في نماذج التعلم الآلي بصريًا. كجزء من أدوات TensorBoard مفتوحة المصدر، يمكن لأدوات What-If تحليل

مجموعات البيانات لتوفر فهمًا لكيفية عمل نماذج التعلم الآلي في ظل سيناريوهات مختلفة، وبناء تصورات خصبة لشرح أداء النموذج. كما تسمح أداة What-If للمستخدم بتعديل عينات من مجموعة البيانات يدويًا ودراسة تأثير هذه التغييرات بواسطة نموذج التعلم الآلي المصاحب. أيضًا، يمكن من خلال تحليل الإنصاف الخوارزمي الموجود بالأداة الكشف عن أنماط التحيز التي لم يكن من الممكن تحديدها من قبل. تسهل أداة What-If على جميع المستخدمين، وإن لم يكونوا مبرمجين، استكشاف واختبار نماذج التعلم الآلي ومعالجة مشكلاتها باستخدام واجهة مستخدم رسومية واضحة وبسيطة.

- Skater: مبادرة أطلقتها Oracle، وهي عبارة عن مكتبة Python لإزالة الغموض عن نموذج الصندوق الأسود، وهي نماذج تنشأ مباشرة من البيانات وبواسطة الخوارزميات، وبالتالي لا يمكن معرفة الكيفية التي يتم بها دمج المتغيرات لصنع التنبؤات. يساعد Skater على بناء نظام تعلم آلي قابل للتفسير يمكن استخدامه على أرض الواقع.