

تحيزات الذكاء الاصطناعي

دليل مرجعي



النسخة الأولى
مايو 2026

المحتويات

08	المقدمة
10	تحيز اجتماعي اقتصادي
11	تحيز إدراكي
12	تحيز استطلاع الرأي
13	تحيز افتراضي
14	تحيز الأتمتة
15	تحيز الاختيار
16	تحيز الأداء
17	تحيز الإرباك
18	تحيز الارتباط
19	تحيز الاستبعاد
20	تحيز الاستدلال البارادوكسي لسيمبسون
21	تحيز الاستقصاء
22	تحيز الامتثال
23	تحيز الانسحاب
24	تحيز الانهيار في السباقات
25	تحيز الأولوية
26	تحيز الباحث
27	تحيز البقاء
28	تحيز البلاغة
29	تحيز التأكيد
30	تحيز التباين
31	تحيز التجانس الخارجي
32	تحيز التجميع
33	تحيز التحول المتكرر في المفهوم
34	تحيز التحول المفاجئ في المفهوم
35	تحيز التخصيص
36	تحيز التخوف
37	تحيز التذكر
38	تحيز الترابط الوهمي

39	تحيز التساهل
40	تحيز التسلسل الزمني
41	تحيز التسمية
42	تحيز التصادم
43	تحيز التصنيف الخاطئ
44	تحيز التطوع
45	تحيز التغذية الراجعة
46	تحيز التغطية
47	تحيز التقرير
48	تحيز التقييم
49	تحيز التوفر
50	تحيز الثقة المفرطة
51	تحيز الجنس
52	تحيز الحداثة
53	تحيز الخوارزمية
54	تحيز الرغبة الاجتماعية
55	تحيز الضبط غير المتزامن
56	تحيز العشوائية
57	تحيز العمر
58	تحيز العينة
59	تحيز القياس
60	تحيز الكشف
61	تحيز المراجع المتباينة
62	تحيز المراجعة
63	تحيز المراقب
64	تحيز المركزية
65	تحيز النطاق
66	تحيز الوصول التشخيصي
67	تحيز انجراف المفهوم
68	تحيز انجراف النموذج
69	تحيز انهيار النموذج
70	تحيز بيئي

71	تحيز تأثير الإجماع الوهمي
72	تحيز تأثير التأطير
73	تحيز تأثير القرون أو تأثير السلبيات
74	تحيز تأثير الهالة
75	تحيز تأثير هوثورن
76	تحيز تاريخي
77	تحيز تأسيسي
78	تحيز تفاعلي
79	تحيز توليدي
80	تحيز حجم العينة
81	تحيز دلالي
82	تحيز رقمي
83	تحيز زمني
84	تحيز سياقي
85	تحيز ضد الأمراض غير المقبولة
86	تحيز ضمني
87	تحيز عدم الاستجابة
88	تحيز عدم القبول
89	تحيز عدم المتانة
90	تحيز عدم المخاطرة
91	تحيز عنصري
92	تحيز غياب التعمية
93	تحيز فرط التخصيص
94	تحيز فرط التعميم
95	تحيز فصلي
96	تحيز استخراج البيانات
97	تحيز الرقابة الأخلاقية
98	تحيز تسرب البيانات
99	تحيز لعبة الاستبدال
100	تحيز لغوي
101	تحيز مرشح الإحالة
102	تحيز معدل القبول

103	تحيز معرفي
104	تحيز نتائج التقرير
105	تحيز نحو الشعبية
106	تحيز نحو المركز
107	تحيز نحو المستجدات
108	تحيز نحو النتائج الإيجابية
109	تحيز وقت الاستجابة
110	تحيز وقت البدء

المقدمة

شهدت تقنيات الذكاء الاصطناعي توسعاً ملحوظاً في مختلف القطاعات الحيوية والحساسة، مثل: العدل، والصحة، والتعليم. ومع هذا التوسع المتسارع، تزايد القلق بشأن تسرب التحيز إلى أنظمة الذكاء الاصطناعي، خاصة تلك التي تؤثر في عمليات صنع القرار، إذ إن الطبيعة المعقدة لهذه الأنظمة والتي تمر بمراحل متعددة تتداخل فيها أدوار عدد من المختصين والخبراء من خلفيات متنوعة، تزيد من احتمالية وجود التحيزات أثناء مراحل التصميم أو التطوير. وقد تؤدي هذه التحيزات إلى تقويض فاعلية أنظمة الذكاء الاصطناعي، حيث قد تتحول من أداة لتعزيز العدالة والإنصاف إلى وسيلة تُفاقم التحيزات بدلاً من الحد منها، مما قد يؤثر سلباً في سمعة المؤسسات، والتعرض لمساءلات قانونية أو شكاوى المستهلكين. ولذلك، فإن حصر هذه التحيزات، وفهم أسبابها، ووضع آليات للحد منها، يمثل خطوة أساسية لضمان نجاح مشاريع الذكاء الاصطناعي، لا سيما في المجالات التي تعتمد على اتخاذ قرارات مؤثرة في الأفراد والمجتمعات.

وفي هذا الإطار، برزت جهود الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا) في تبني منهج واضح وشامل لتعزيز عدالة أنظمة الذكاء الاصطناعي والحد من التحيز فيها. فقد أصدرت الهيئة "مبادئ أخلاقيات الذكاء الاصطناعي" التي تؤكد العدالة وعدم التمييز، وعززت ذلك بإطلاق "مبادئ الذكاء الاصطناعي التوليدي للجهات الحكومية" و"مبادئ الذكاء الاصطناعي التوليدي للعموم" بما يساهم في توجيه الاستخدام المسؤول لهذه التقنيات. ولم تقتصر جهودها على الجانب الإرشادي، بل امتدت إلى تطوير "إطار تبني الذكاء الاصطناعي" ليكون مرجعاً دائماً للجهات في تبني التقنيات بصورة منهجية ومسؤولة، عبر تعزيز الحوكمة المؤسسية وبناء القدرات. وعلى الصعيد المعرفي، نشرت سدايا دراسة بعنوان "التحيز في أنظمة الذكاء الاصطناعي: تحديات وحلول"، استعرضت فيها مواضيع ظهور التحيز عبر مراحل بناء الأنظمة، وقدمت أدوات وإرشادات عملية تساهم في اكتشافه والحد من آثاره.

وبناءً على ذلك، يأتي هذا الدليل ليقدّم أبرز أنواع التحيزات (أكثر من مائة تحيز) التي قد تؤثر في دقة وعدالة أنظمة الذكاء الاصطناعي، ويسلّط الضوء على تعريفها وكيفية نشوئها وتأثيرها في المجتمع، مع استعراض أمثلة واقعية واستراتيجيات فاعلة للتخفيف من آثارها، وصولاً إلى تطوير أنظمة ذكاء اصطناعي أكثر شفافية وإنصافاً.



تحيز اجتماعي اقتصادي

يحدث عندما تطور أو تعزز أنظمة الذكاء الاصطناعي عدم المساواة بناءً على الوضع الاقتصادي للأفراد، أو التعليم، أو المهنة، أو غيرها من العوامل الاجتماعية والاقتصادية. وينشأ هذا التحيز عادة من مجموعات البيانات التي لا تمثل بشكل كافٍ بعض المجموعات الاجتماعية والاقتصادية أو من الخوارزميات التي تفضل عن غير قصد الخصائص الأكثر شيوعاً بين الأشخاص ذوي الوضع الاجتماعي والاقتصادي الأعلى.

يؤدي التحيز الاجتماعي الاقتصادي إلى تعميق الفجوات القائمة من خلال حرمان الأفراد من الخلفيات المنخفضة الدخل أو التعليم من فرص متكافئة في الوصول إلى الموارد أو الحصول على معاملة عادلة. كما يسهم هذا التحيز في استبعاد احتياجات الفئات المهمشة أو إساءة فهم ظروفها، مما يرسخ التمييز داخل تطبيقات الذكاء الاصطناعي.



التأثير

قد تُظهر أدوات التوظيف المعتمدة على الذكاء الاصطناعي تحيزاً اجتماعياً واقتصادياً عند تفضيل المرشحين من خلفيات تعليمية نخوية، مما يمنح الأفضلية بشكل غير عادل لمن تتاح لهم هذه الفرص بسبب أوضاعهم المادية. يؤدي ذلك إلى إقصاء مرشحين مؤهلين من خلفيات أقل حظاً، ويحد من تنوع القوى العاملة وعدالة الفرص.



المثال

تنوع مجموعات بيانات التدريب لضمان تمثيل مجموعة واسعة من الخلفيات الاجتماعية والاقتصادية. كما يجب على المطورين إجراء فحص ناقذ وتعديل المعايير والميزات التي تستخدمها نماذج الذكاء الاصطناعي لمنع التفضيل غير المقصود لبعض المؤشرات الاجتماعية والاقتصادية. بالإضافة إلى ذلك، يمكن أن يساعد تطبيق تقنيات تعلم الآلة المعنية بالإنصاف -مثل: قيود العدالة الخوارزمية وأساليب تصحيح التحيز- على تقليل الفوارق الاجتماعية والاقتصادية.



التخفيف

تحيز إدراكي

يشير إلى التحيز الذي ينشأ نتيجة الافتراضات أو التقديرات المسبقة المضمّنة في تصميم أنظمة الذكاء الاصطناعي، والتي تؤدي إلى قرارات غير عادلة تجاه فئات معينة، رغم عدم وجود مبررات موضوعية كافية. وغالباً ما يُعبر هذا التحيز عن وجهات نظر أو تصورات ضمنية لدى مطوري النظام أو في كيفية تفسير البيانات.

يؤثر التحيز الإدراكي في دقة وعدالة أنظمة الذكاء الاصطناعي من خلال ترسيخ تفضيلات أو افتراضات ضمنية ناتجة عن البيانات أو تصميمات الخوارزميات. يؤدي هذا التحيز إلى نتائج تميل إلى صالح مجموعات معينة دون مبرر موضوعي، مما يضعف مصداقية النظام ويهدد بترسيخ أحكام غير عادلة في القرارات الآلية.



في نظام تقدير الائتمان للقروض، يعتمد النظام على بيانات التاريخ الائتماني والدخل لاتخاذ قرارات بشأن منح القروض. ومع ذلك، يُظهر النظام تحيزاً حيث يرفض بشكل أكبر طلبات القروض من الشباب ذوي تاريخ ائتماني قصير أو دخل محدود، مقارنة بالفئات العمرية أو الاجتماعية الأخرى.



اختيار بيانات تدريب متنوعة وشاملة، وفحص وتنظيف البيانات للتعرف على أي أنماط تحيز وإزالتها. كذلك، ينبغي تطوير خوارزميات واعية بالتحيز يمكنها التعرف عليه وتقليله، وإجراء اختبارات شاملة للنماذج على مجموعات بيانات متعددة لضمان أداء عادل. وأخيراً، يجب القيام بتدقيق ومراجعة مستمرة من قبل فرق متعددة التخصصات لضمان الشفافية والعدالة في النتائج.



تحيز استطلاع الرأي

يشير إلى عدم الدقة في نماذج الذكاء الاصطناعي التي تنتج بالاعتماد على البيانات الناتجة عن المعلومات المبلغ عنها ذاتياً. يحدث هذا النوع من التحيز عندما يقدم الأفراد إجابات ذاتية أو غير دقيقة أو غير كاملة في الاستطلاعات التي تعمل بعد ذلك كبيانات أساسية لتدريب أنظمة الذكاء الاصطناعي. ينبع هذا التحيز من التناقضات بين السلوكيات أو الخصائص أو التفضيلات المبلغ عنها وحالاتها الفعلية، مما يؤدي إلى انحراف مدخلات البيانات لنماذج الذكاء الاصطناعي.

يؤدي تحيز استطلاع الرأي إلى تطوير نماذج ذكاء اصطناعي تستند إلى ارتباطات مضللة نتيجة الاعتماد على بيانات مُبلغ عنها ذاتياً، قد تعكس انحيازات أو تصورات غير دقيقة. يُضعف هذا التحيز دقة التنبؤات ويقود إلى قرارات غير عادلة، مما يؤثر سلباً في موثوقية وفعالية الأنظمة المعتمدة على تلك البيانات.



التأثير

نموذج الذكاء الاصطناعي المصمم للتنبؤ بمخاطر الصحة الغذائية بناءً على استطلاعات استهلاك الغذاء المبلغ عنها ذاتياً قد يكون متحيزاً من قبل المشاركين الذين لا يبلغون عن عادات الأكل غير الصحية أو يقومون بالإبلاغ عن العادات الصحية بشكل كبير، ويمكن أن يؤدي هذا الاختلاف بين الأنماط الغذائية المبلغ عنها والفعلية إلى تقييمات غير دقيقة للمخاطر.



المثال

إضافة أساليب للتحقق من البيانات المُبلغ عنها ذاتياً وتصحيحها حيثما أمكن ذلك، باستخدام مصادر البيانات الموضوعية كمكمل أو موازنة للمعلومات المبلغ عنها ذاتياً. يمكن أن يشمل ذلك إجابة استجابات المسح إلى البيانات السلوكية، وتنفيذ تقنيات التحقق من صحة البيانات أثناء مرحلة تصميم الاستطلاع لتقليل احتمالية الإبلاغ غير الدقيق، واستخدام النماذج الإحصائية لضبط التحيزات المعروفة في الإبلاغ الذاتي.



التخفيف

تحيز افتراضي

يشير إلى الانحرافات التي قد تظهر في أداء نماذج الذكاء الاصطناعي نتيجة الاعتماد على بيانات محاكاة أو بيانات افتراضية عند التدريب أو التقييم، بدلاً من بيانات واقعية. هذا النوع من التحيز الافتراضي قد يؤدي إلى نتائج غير دقيقة أو غير عادلة عند تطبيق النموذج في سياقات الحياة الحقيقية التي تختلف عن ظروف المحاكاة.

يؤثر التحيز الافتراضي في موثوقية نماذج الذكاء الاصطناعي عند تطبيقها في سياقات واقعية، مما يؤدي إلى قرارات غير دقيقة أو غير عادلة. يظهر هذا الأثر بشكل خاص في المجالات الحساسة مثل التوظيف أو الإقراض أو العدالة الجنائية، حيث يمكن أن تتسبب الأنظمة المتحيزة في الإضرار بفئات معينة أو ترسيخ التحيزات المجتمعية القائمة.



التأثير

نظام الموافقة على القروض مدرب على مجموعة بيانات من المقترضين الافتراضيين ذوي الخصائص والتاريخ الائتماني الموحدة. من الممكن أن يؤدي النموذج أداءً جيداً على هذه البيانات المجمعة، إلا أنه قد يفشل في التقييم الدقيق للجدارة الائتمانية للمقترضين في العالم الواقعي من ذوي الخلفيات والأوضاع المالية المتنوعة، مما قد يؤدي إلى ممارسات إقراض غير عادلة.



المثال

استخدام بيانات تمثيلية ومتنوعة من العالم الحقيقي لتدريب واختبار نماذج الذكاء الاصطناعي، مما يضمن تعرضها للتعقيد والتنوع الفعلي في البيانات والسياقات الواقعية. هذا النهج يقلل من اعتماد النماذج على الافتراضات غير الدقيقة التي قد تنتج عن بيانات المحاكاة، ويعزز دقة الأداء وعدالة النتائج في التطبيقات العملية.



التخفيف

تحيز الأتمتة

يشير إلى ميل البشر لتفضيل القرارات، أو التنبؤات، أو النتائج المولدة بواسطة أنظمة الآلة على تلك المستنتجة من الحكم البشري أو مصادر أخرى. يمكن أن يحدث هذا التحيز بين المطورين، أو المستخدمين، أو أصحاب المصلحة الذين قد يعتمدون بشكل زائد على أنظمة الذكاء الاصطناعي، بفرض أن قرارات الآلة أكثر موثوقية أو موضوعية من تلك التي يتخذها البشر، حتى عندما تشير الأدلة إلى خلاف ذلك.

يؤثر هذا التحيز في عدد من المجالات الحيوية من خلال تعزيز الثقة المفرطة في الأنظمة الآلية، مما يقلل من الانتباه إلى أخطائها أو قيودها. في البيئات الحساسة، يمكن أن تقود هذه الثقة غير المبررة إلى تجاهل الخبرة البشرية، أو الفشل في اكتشاف العيوب، أو التغاضي عن رؤى دقيقة قد تكون حاسمة في اتخاذ القرار.



التأثير

نظام ذكاء اصطناعي يُستخدم لتحديد أولوية الحالات في أقسام الطوارئ، ويحظى بثقة عالية من الطاقم الطبي، مما يدفعهم إلى اتباع توصياته دون مراجعة. يؤدي ذلك إلى التقليل من أهمية بعض الأعراض لدى فئات عمرية معينة، ويُنتج قرارات قد تؤخر التدخل الطبي اللازم للحالات الحرجة.



المثال

اعتماد اختبارات شاملة وتقنيات تحقق دقيقة، إلى جانب تطبيق آليات مراجعة وتغذية راجعة منتظمة خلال دورة تطوير النظام. كما يُنصح بتجنب استخدام بيانات ينتجها الذكاء الاصطناعي نفسه، نظراً إلى احتمالية تعزيزها لتحيزات قائمة، وعجزها عن تمثيل تعقيد وتنوع البيانات الحقيقية بشكل كافٍ.



التخفيف

تحيز الاختيار

يحدث عندما لا تمثل البيانات المدخلة لنظام الذكاء الاصطناعي تمثيلاً متوازناً لنطاق كامل السيناريوهات ذات الصلة، مما يؤدي إلى نتائج منحرفة أو متحيزة. وتحيز الاختيار هو من أنواع تحيز العينة ويمكن أن ينجم عن طريق جمع البيانات غير العشوائية أو استبعاد مجموعات معينة أو تمثيل زائد لأنواع معينة من البيانات.

يؤدي تحيز الاختيار إلى تفاوت في جودة الخدمات المقدمة بين فئات مختلفة من المستخدمين. كما يساهم هذا التحيز في تعزيز الفوارق الاجتماعية أو الاقتصادية القائمة، وقد يؤدي إلى تفاقمها بمرور الوقت نتيجة استبعاد منهجي لفئات معينة من الاعتبارات الخوارزمية.



التأثير

في أدوات التوظيف المعتمدة على الذكاء الاصطناعي، إذا كانت مجموعة البيانات المستخدمة لتدريب الذكاء الاصطناعي تتضمن أغلبية السير الذاتية للمرشحين الذكور، وخاصة في الأدوار الفنية، فقد يطور النظام تحيزاً لصالح المرشحين الذكور للمناصب المستقبلية، مما يقلل من فرص المرشحات الإناث المؤهلات.



المثال

ضمان شمولية مجموعات البيانات المستخدمة في مراحل التدريب والاختبار والتحقق، بحيث تمثل جميع فئات المستخدمين والسيناريوهات المحتملة بشكل متوازن. ويسهم هذا التنوع في تحسين قدرة النموذج على التعميم وتقليل احتمالية التحيز ضد فئات أو حالات غير ممثلة بشكل كافٍ في البيانات.



التخفيف

تحيز الأداء

يشير هذا التحيز إلى حالة يُظهر فيها النموذج أداءً منخفضاً على أجزاء أو مجموعات فرعية معينة من البيانات، على الرغم من أدائه المرضي على مجموعة البيانات الكاملة.

يؤثر تحيز الأداء في أنظمة الذكاء الاصطناعي من خلال تفاوت دقة النتائج بين مجموعات مختلفة من المستخدمين أو الحالات، وينتج هذا التحيز عن عدم توازن في أداء النموذج، حيث يعمل بكفاءة أعلى لفئات معينة مقارنة بآخرى، مما يؤدي إلى قرارات غير متكافئة قد تُضعف العدالة والموثوقية في التطبيقات الحساسة.



التأثير

نظام للتعرف على الوجوه يُستخدم في سياقات أمنية يتم تدريبه بشكل أساسي على صور لأفراد من خلفية عرقية محددة، مما يؤدي إلى انخفاض كبير في دقته عند التعامل مع وجوه من مجموعات عرقية أخرى وينتج عن هذا الخلل في الأداء ارتفاع النتائج الإيجابية الخاطئة، مما يؤدي إلى تحيز ضد أفراد لم يتم تمثيلهم بشكل كافٍ في بيانات التدريب.



المثال

استخدام بيانات تدريب متوازنة تمثل التنوع داخل المجموعة المستهدفة، إلى جانب تطبيق تدخلات خوارزمية تعزز العدالة أثناء تطوير النموذج. ويسهم هذا النهج في ضبط الفوارق في الأداء بين الفئات المختلفة، ويعزز من دقة النظام وموثوقيته عبر جميع الحالات المحتملة.



التخفيف

تحيز الإرباك

ينشأ تحيز الإرباك عندما يفشل نموذج الذكاء الاصطناعي في تحديد العلاقة الحقيقية بين المتغيرات مما يؤدي إلى انحراف في استنتاجات نموذج الذكاء الاصطناعي، وينشأ هذا التحيز عندما يفشل النموذج في أخذ متغير ثالث بعين الاعتبار، يكون له تأثير في كل من المتغير التنبؤي والمتغير المستهدف، مما يؤدي إلى استنتاجات غير دقيقة أو مضللة في المخرجات.

يؤدي تحيز الإرباك إلى اعتماد نماذج الذكاء الاصطناعي على علاقات غير دقيقة بين المتغيرات، نتيجة تجاهل عوامل مشوشة تؤثر في التفسير السببي. يُقوض هذا التحيز موثوقية ودقة التنبؤات، وقد ينتج عنه قرارات غير فعالة أو حتى ضارة، خاصة في المجالات التي تتطلب فهماً دقيقاً للعلاقات السببية مثل الرعاية الصحية أو السياسات العامة.



التأثير

محاولة التنبؤ بمبيعات المثلجات بناءً على درجة الحرارة، وتجاهل حقيقة أن المبيعات تتأثر أيضاً أثناء أيام الأسبوع، فترتفع نسبتها في عطلة نهاية الأسبوع، بغض النظر عن درجة الحرارة. فإذا أخذت في الاعتبار درجة الحرارة فقط، سينسب نموذج الذكاء الاصطناعي الخاص بك بشكل غير صحيح الزيادة في المبيعات استناداً إلى درجة الحرارة فقط دون ملاحظة التأثير الذي قد يحدث أثناء أيام الأسبوع.



المثال

دراسة جميع المتغيرات ذات الصلة التي قد تؤثر في النتائج بعناية وأخذها في الاعتبار، مع التركيز على تحديد الإرباكات المحتملة ودمجها ضمن النموذج. كما يساهم التقييم المنتظم لأداء النموذج وتحليل الحساسية في ضمان التقاط الآليات السببية الأساسية بدقة دون التأثير بالمتغيرات المربكة.



التخفيف

تحيز الارتباط

يحدث عندما لا تعكس نماذج الذكاء الاصطناعي التحيزات الموجودة فحسب، بل تعمل على تضخيمها من خلال عملية التدريب، ويتجذر التحيز في الارتباطات التي تعلمتها من مجموعة بيانات التدريب التي قد تحتوي على قرارات تاريخية متحيزة، أو صور نمطية اجتماعية، أو تمثيلات غير متوازنة.

يؤثر تحيز الارتباط بشكل كبير في الأفراد والمجتمعات من خلال تعزيز الصور النمطية الضارة واستمرار أنماط عدم المساواة. لا يقتصر الأثر على الفئات المتأثرة مباشرة، بل يمتد ليقوض نزاهة أنظمة الذكاء الاصطناعي ويُضعف ثقة المستخدمين في عدالة مخرجاتها على نطاق أوسع.



التأثير

خوارزمية توقع جرائم المخدرات تُدرَّب على بيانات متحيزة ناتجة عن الفصل السكني والممارسات الشرطية غير المتوازنة، مما يؤدي إلى توجيه الشرطة بشكل متكرر نحو أحياء تضم أقليات سكانية. تُنتج هذه الممارسات مزيداً من الاعتقالات في تلك المناطق، وتُعاد تغذية البيانات الجديدة إلى النظام، مما يعزّز دورة مستمرة من التحيز والتطبيق غير العادل.



المثال

اتباع نهج شامل في مراحل مختلفة من التطوير والنشر، بحيث يتضمن ذلك ضمان بيانات تدريب متنوعة ومتوازنة تمثل بدقة السكان المستهدفين، واستخدام تقنيات الكشف عن التحيز لتحديد التحيزات المحتملة، وتنفيذ خوارزميات مدركة للعدالة لتقليل التحيز في التنبؤات، والمراقبة والتقييم المستمر لأداء النظام بعد النشر لمعالجة أي تحيزات ناشئة.



التخفيف

تحيز الاستبعاد

يحدث تحيز الاستبعاد عندما ينحرف أداء نظام الذكاء الاصطناعي نتيجة حذف معلومات مهمة من مجموعة البيانات المُستخدمة في التدريب أو التقييم، وغالباً ما ينشأ هذا التحيز بسبب تجاهل عوامل ذات صلة أو التقليل من أهمية سمات معينة، مما يؤدي إلى نماذج غير مكتملة لا تعكس الصورة الحقيقية للمجال المستهدف.

يؤدي تحيز الاستبعاد إلى أنظمة ذكاء اصطناعي غير مطلعة بشكل كافٍ وقد تكون متحيزة ضد مجموعات أو سيناريوهات معينة غير ممثلة في بيانات التدريب. ويمكن أن يؤثر هذا في فاعلية النموذج وعدالته، مما ينتج عنه تنبؤات أو توصيات أو قرارات منحازة. ونتيجة لذلك، قد يطور نموذج الذكاء الاصطناعي نقاطاً عمياء، حيث يفتقر إلى النطاق الكامل من المعلومات اللازمة لاتخاذ قرارات دقيقة وعادلة.



التأثير

نظام الذكاء الاصطناعي لفحص طلبات الوظائف وتحديد المرشحين الأكثر تأهيلاً. يتم تدريب نظام الذكاء الاصطناعي على مجموعة بيانات للتعيينات الناجحة السابقة في الشركة، لكن مجموعة البيانات هذه لا تتضمن سوى معلومات عن المرشحين الذين لديهم مؤهلات تعليمية وخبرة عملية معينة. يمكن أن يؤدي تحيز الاستبعاد هذا إلى إغفال نظام الذكاء الاصطناعي المرشحين الذين لديهم خلفيات تعليمية أو عملية مختلفة غير ممثلة في مجموعة البيانات، ونتيجة لذلك، قد تفوت الشركة توظيف مواهب متنوعة.



المثال

تضمين إجراء مراجعات شاملة لعمليات جمع البيانات وإعدادها لضمان التغطية الشاملة لجميع العوامل والمجموعات ذات الصلة. قد يشمل ذلك البحث عن بيانات المجموعات السكانية الممثلة تمثيلاً ناقصاً أو المتغيرات التي تم تجاهلها سابقاً ودمجها. يمكن أن يساعد استخدام تقنيات -مثل: تحليل أهمية الميزة- على تحديد المعلومات المفقودة وتصحيحها.



التخفيف

تحيز الاستدلال البارادوكسي لسيمبسون

يحدث تحيز الاستدلال البارادوكسي لسيمبسون عندما يظهر نمط أو استنتاج معين عند تحليل البيانات ضمن مجموعات فرعية، لكنه يختفي أو ينقلب عند دمج البيانات ككل. يعكس هذا التحيز مفارقة إحصائية تؤدي إلى استنتاجات متضاربة بناءً على طريقة تنظيم البيانات، مما قد يُضلل تفسير نتائج النموذج.

يؤدي هذا التحيز إلى استنتاجات مضللة في أنظمة الذكاء الاصطناعي عند الاعتماد على بيانات مجمعة تخفي العلاقات الفعلية بين المتغيرات. يؤثر ذلك سلباً في دقة النماذج، خاصة في حالات اتخاذ القرار المعتمدة على التحليل الإحصائي، مما قد ينتج عنه قرارات خاطئة أو مخرجات متحيزة إذا لم يتم التعرف على المفارقة ومعالجتها.



التأثير

برنامج تعليمي جديد يهدف إلى تحسين أداء الطلاب، عند تحليل البيانات بشكل منفصل لمدرستين، المدرسة "أ" والمدرسة "ب"، قد يبدو البرنامج أكثر فاعلية في المدرسة "أ"، وعندما يتم دمج البيانات الخاصة بكليهما المدرستين قد ينعكس الاتجاه، مما يظهر أن البرنامج أكثر فاعلية بشكل عام في المدرسة "ب"، ويمكن أن يحدث هذا الانعكاس بسبب الاختلافات في الحجم أو التركيبة السكانية. وعند عدم التعرف على هذه الاختلافات وتفسيرها يمكن لنظام الذكاء الاصطناعي أن يتوصل إلى استنتاجات غير صحيحة.



المثال

تطبيق استراتيجيات مثل التحليل الطبقي لفهم اتجاهات المجموعات الفرعية قبل تجميع البيانات، ودمج المعلومات السياقية في نماذج الذكاء الاصطناعي لمعرفة متى قد تكون البيانات المجموعة مضللة، وإنشاء عمليات مراجعة منتظمة والتحقق من الصحة لضمان بقاء مخرجات النموذج دقيقة وتعكس الواقع.



التخفيف

تحيز الاستقصاء

ينشأ تحيز الاستقصاء عندما تحدث أخطاء منهجية في جمع البيانات أو معالجتها داخل أنظمة الذكاء الاصطناعي، وينشأ هذا التحيز عند استخدام طرق غير دقيقة أو غير كافية في جمع البيانات أو عند تصنيفها بطريقة لا تعكس الواقع، مما يؤدي إلى تشويه تمثيل الظواهر المدروسة في نتائج النموذج.

يؤدي تحيز الاستقصاء إلى إنشاء نماذج لا تعكس بدقة خصائص السكان أو الظواهر المستهدفة مما يُنتج تنبؤات أو قرارات متحيزة. يمتد تأثير هذا التحيز إلى مجالات متعددة كالرعاية الصحية والتسويق، وقد يؤدي إلى استنتاجات خاطئة. كما ينشأ غالباً من خلل في جمع البيانات أو تسمية العينات أو نقص تمثيل مجموعات محددة.



نموذج ذكاء اصطناعي للتعرف على المشاعر تم تدريبه على بيانات مأخوذة من شبكات اجتماعية يستخدمها الشباب، مما أدى إلى ضعف دقته عند تفسير نصوص كتبها كبار السن أو فئات أخرى.



التأكد من أن بيانات التدريب متنوعة وممثلة لحالات الاستخدام المقصودة، مع إجراء عمليات تدقيق منتظمة لكل من البيانات وأداء النماذج، واستخدام تقنيات متقدمة لاكتشاف التحيز ومعالجته بهدف تعزيز موثوقية النتائج وتحسين أداء النماذج في مختلف السيناريوهات المتوقعة.



تحيز الامتثال

يشير إلى الحالة التي يتم فيها تدريب أو تصميم أنظمة الذكاء الاصطناعي لتعطي الأولوية للنتائج التي تتماشى مع المتطلبات التنظيمية أو القانونية، أو سياسات المؤسسة، أو الأعراف الاجتماعية، بدلاً من التحسين من أجل الدقة، أو العدالة، أو معايير أخرى ذات صلة. يمكن أن يحدث هذا التحيز عندما يعطي مطورو الذكاء الاصطناعي أو المنظمات الأولوية للالتزام بالقواعد أو الإرشادات على تحقيق أفضل النتائج الممكنة لجميع أصحاب المصلحة.

يؤثر تحيز الامتثال في أداء أنظمة الذكاء الاصطناعي من خلال تركيزها المفرط على الالتزام بالقواعد التنظيمية، على حساب أهداف جوهرية مثل العدالة أو رضا المستخدم، وقد يؤدي هذا التحيز إلى تطوير أنظمة أقل ابتكاراً، وتهميش الاعتبارات الأخلاقية، مما يُضعف الثقة في التطبيقات الذكية ويحد من قدرتها على تلبية الاحتياجات الفعلية للمجتمع.



التأثير

نظام الذكاء الاصطناعي لفحص السير الذاتية للمتقدمين على الوظائف، وللامتثال لسياسات التنوع، يُطلب من الذكاء الاصطناعي إعطاء الأولوية للمرشحين من المجموعات الأقل تمثيلاً. ولكن الذكاء الاصطناعي يركز بشكل مفرط على أساس الجنس أو العرق فقط ويؤدي ذلك بشكل غير عادل إلى رفض المرشحين المؤهلين من المجموعات الممثلة تمثيلاً زائداً.



المثال

تحسين خوارزميات الذكاء الاصطناعي للنظر في مجموعة واسعة من العوامل على دمج المبادئ الأخلاقية والعدالة في التصميم الخوارزمي منذ البداية، مما يضمن أن القرارات ليست مدفوعة بالامتثال فحسب، بل أيضاً بالتأثير الأوسع في جميع أصحاب المصلحة. يمكن أن يشمل ذلك دمج آليات التغذية الراجعة التي تسمح بالتعديل الديناميكي للنموذج بناءً على نتائج العالم الحقيقي.



التخفيف

تحيز الانسحاب

يحدث تحيز الانسحاب عندما يتوقف بعض المستخدمين عن التفاعل مع النظام، غالباً بسبب إخفاق النموذج في تلبية توقعاتهم. يؤثر هذا التحيز في جودة البيانات المتبقية، مما يؤدي إلى نتائج غير موثوقة وتمثيل غير دقيق، ويقلل من قدرة النظام على العمل بشكل عادل عبر جميع فئات المستخدمين.

يؤدي إلى تدهور جودة وموثوقية تنبؤات نظام الذكاء الاصطناعي بمرور الوقت، حيث تصبح البيانات غير ممثلة لقاعدة المستخدمين الأوسع أو السكان المستهدفين. يؤدي هذا التحيز إلى تحريف عملية التعلم الخاصة بنظام الذكاء الاصطناعي، مما قد يتسبب في ضعف أدائه لمجموعة واسعة من المستخدمين، ونتيجة لذلك قد يفشل النظام في معالجة أو حتى الاعتراف باحتياجات وتفضيلات أولئك الذين توقفوا عن استخدامه، مما يخلق دورة تزيد من ترسيخ قيود النظام وتحيزاته.



نظام توصية في تطبيق لعرض الأفلام يعتمد على تفضيلات المشاهدين النشطين، مع مرور الوقت، يتوقف بعض المستخدمين الذين يفضلون أفلاماً أقل شهرة أو أنماطاً فنية متنوعة عن استخدام التطبيق بسبب التوصيات غير الملائمة. نتيجة لذلك، يصبح النظام أكثر انحيازاً نحو الأنواع الشائعة التي يفضلها المستخدمون الذين بقوا، مما يؤدي إلى تضيق نطاق التوصيات وإهمال التنوع في أذواق المشاهدين الآخرين.



أخذ المستخدمين الذين توقفوا عن استخدام النظام في الاعتبار، والسعي إلى تحقيق تمثيل أكثر تنوعاً للمستخدمين، وضبط الخوارزميات باستمرار لتكون أكثر عمومية وتجنب الإفراط في التوافق مع تفضيلات المستخدمين الذين يظلون متفاعلين من خلال تغذية البيانات المتنوعة.



تحيز الانهيار في السياقات

يحدث تحيز الانهيار في السياقات عندما تفشل أنظمة الذكاء الاصطناعي في فهم أو دمج التفاصيل السياقية الدقيقة المرتبطة بالتفاعلات البشرية، مما يمثل قصوراً في قدرة أنظمة الذكاء الاصطناعي. يشمل هذا القصور عوامل مثل اللهجات، والإشارات الثقافية، والديناميكيات الاجتماعية، والنوايا الضمنية، مما يؤثر في دقة تفسير المحتوى والاستجابة له في السياقات المتنوعة.

يؤدي تحيز الانهيار في السياقات إلى أن أنظمة الذكاء الاصطناعي تفسر المعلومات بشكل غير صحيح أو تبسطها بشكل مفرط، مما يؤدي إلى سوء الفهم، وتشويه الصورة، وحتى تفاقم النزاعات في البيئات الرقمية. يمكن أن يؤدي هذا إلى تقويض فاعلية أدوات التواصل التي تعتمد على الذكاء الاصطناعي، ويسهم في انتشار الأخبار الكاذبة، وتعزيز التحيزات أو الصور النمطية القائمة بالفعل.



التأثير

عندما تقوم أنظمة إدارة المحتوى المدعومة بالذكاء الاصطناعي في منصات التواصل الاجتماعية، بوضع علامة على المحتوى أو إزالته بشكل خاطئ بسبب عدم فهم السياق الأساسي. على سبيل المثال: قد تتم إساءة تفسير نكتة أو نقد ساخر غير ضار على أنه كلام مسيء أو ضار، مما يؤدي إلى حجب غير مبرر أو قمع في التعبير.



المثال

تكييف نماذج الذكاء الاصطناعي مع الظروف والفروق الدقيقة المحددة لكل سياق جديد في المكان الذي سيتم تطبيقها فيه. ويشمل ذلك إعادة تدريب النماذج باستخدام البيانات المحلية، وفهم الاختلافات الثقافية أو التنظيمية، وضمان تلبية أنظمة الذكاء الاصطناعي لاحتياجات أصحاب المصلحة المحليين. كذلك، إشراك خبراء المجال، وعلماء الأخلاق، والمجتمعات المحلية أثناء مراحل التطوير والنشر لضمان مراعاة نظام الذكاء الاصطناعي للعوامل السياقية الفريدة، بما في ذلك الثقافة والقانون والأعراف الاجتماعية. وإجراء اختبارات مكثفة لأنظمة الذكاء الاصطناعي في بيئاتها التشغيلية التي سيتم تشغيلها فيها.



التخفيف

تحيز الأولوية

يشير إلى ميل الخوارزميات إلى إعطاء أولوية أو وزن غير متناسب لأولى المعلومات التي تعالجها أو للبيانات الأولية في تسلسل، ويمكن أن يعكس هذا التحيز التحيزات المعرفية البشرية، حيث تؤثر المعلومات المبكرة في المعالجة واتخاذ القرارات التالية بشكل أقوى من المعلومات المقدمة لاحقاً.

يؤثر في موضوعية ودقة النتائج عبر التطبيقات المختلفة، بدءاً من محركات البحث وحتى توصية المحتوى. ومن خلال تفضيل المعلومات التي يتم الحصول عليها في وقت مبكر. قد يعزز النظام التحيزات الحالية ويحتمل أن يؤدي إلى تهميش المعلومات الأحدث أو الأقل أهمية، ويمكن أن يحد من تنوع المحتوى أو وجهات النظر المقدمة للمستخدمين.



نظام تصنيف لمقالات الأخبار يمنح وزناً أكبر للمعلومات التي يتم استقبالها أولاً. إذا كانت الفقرات الأولى للمقال تحتوي على تحيزات أو معلومات غير دقيقة، فإن النظام يبني تقييمه الكامل للمقال بناءً على هذه المعلومات المبكرة، مما يؤدي إلى تهميش التصحيحات أو التوضيحات التي قد ترد لاحقاً في النص.



تصميم أنظمة الذكاء الاصطناعي لتقييم المعلومات ونقاط البيانات بشكل أكثر إنصافاً عبر التسلسلات. ويمكن تحقيق ذلك من خلال تنفيذ خوارزميات تعمل على ترتيب عشوائي لترتيب معالجة البيانات بشكل متعمد أو من خلال دمج الآليات التي تتكيف مع المبالغة المحتملة في تقدير البيانات المقدمة مبكراً. يمكن أن يساعد التحديث المنتظم لبيانات تدريب الذكاء الاصطناعي لتشمل مدخلات حديثة ومتنوعة على تقليل تأثيرات التحيز للأولوية.



تحيز الباحث

يحدث تحيز الباحث عندما يؤثر مطور النموذج أو الباحث في مخرجات نظام الذكاء الاصطناعي بشكل غير مقصود، نتيجة محاولته المستمرة لتوجيه النموذج نحو نتائج تتماشى مع فرضياته أو توقعاته المسبقة. يظهر هذا التحيز غالباً أثناء عمليات التدريب أو التقييم، مما يؤدي إلى تشويه حيادية النموذج.

يؤدي إلى المساس بموضوعية وموثوقية وصحة أنظمة الذكاء الاصطناعي وذلك عندما يتم تجهيز النموذج أو تصميمه ليتوافق مع توقعات مطوريه، فقد يؤدي أداءً جيداً في اختبار السيناريوهات التي تتوافق مع هذه التوقعات ولكنها تفشل في تطبيقات العالم الحقيقي. يمكن أن يؤدي هذا التحيز إلى نشر حلول ليست فعالة أو عادلة كما هو مفترض، مما قد يتسبب في اتخاذ قرارات مضللة، أو سوء تخصيص الموارد، أو إلحاق الضرر بالأفراد المتأثرين بهذه القرارات المتحيزة.



نظام ذكاء اصطناعي لتقييم كفاءة الموظفين قام بتطويره فريق يؤمن بأن سنوات الخبرة هي العامل الأهم للنجاح. أثناء بناء النموذج، ركز الباحثون بشكل مفرط على ميزات مرتبطة بعدد سنوات العمل وأعطوها وزناً أعلى، مما جعل النموذج يميل تلقائياً إلى تفضيل المرشحين الأكثر خبرة بغض النظر عن الكفاءة الحقيقية أو الأداء الفعلي.



تطبيق وتنفيذ تقنيات التحليل الأعمى، حيث يُمنع منشئو النماذج من معرفة النتائج المرتبطة بفرضياتهم أثناء مراحل تدريب النماذج وتقييمها، مما يسهم في تقليل احتمالية التوجيه اللاواعي للنتائج وضمان قدر أكبر من الموضوعية والدقة في الأداء النهائي للنماذج.



تحيز البقاء

يحدث تحيز البقاء عندما يتركز تقييم الأداء أو النتائج على الحالات الناجحة، مع تجاهل الإخفاقات أو العينات غير المكتملة. يؤدي هذا التحيز إلى رسم صورة أكثر إيجابية مما يعكسه الواقع فعلياً، مما يُنتج انحرافاً في متوسط النتائج نحو الأعلى ويؤثر في دقة التقييمات والتحليلات.

يؤثر تحيز البقاء في موثوقية مخرجات أنظمة الذكاء الاصطناعي من خلال تشويه التوقعات والمبالغة في تقدير معدلات النجاح، مقابل التقليل من تقدير المخاطر. يُضعف هذا التحيز دقة النماذج في اتخاذ القرار، خاصة في السيناريوهات التي تتطلب فهماً شاملاً لكامل النتائج المحتملة، بما في ذلك الإخفاقات.



التأثير

نموذج تحليل الأداء المالي قد يعاني من تحيز البقاء عندما يستند إلى بيانات الشركات التي ما زالت قائمة فقط، متجاهلاً الشركات التي فشلت وخرجت من السوق. يؤدي هذا التحيز إلى انحراف نتائج التحليل، إذ يتم تقييم النجاح بناءً على عينة غير ممثلة تقتصر على الشركات التي استطاعت الاستمرار.



المثال

إضافة نقاط البيانات بشكل فعال التي تعكس كلاً من النجاحات والإخفاقات في مجموعة البيانات، والتحقق من تنبؤات النموذج مقابل مجموعات البيانات الخارجية أو إجراء تجارب في العالم الحقيقي لاكتشاف الاختلافات، وتحديث النموذج باستمرار بالبيانات المجمعة حديثاً.



التخفيف

تحيز البلاغة

يشير إلى الطرق التي قد تؤدي بها استراتيجيات اللغة والتواصل إلى إدخال أو استمرار التحيزات في المناقشات أو الأوضاع الاقتصادية أو السرديات حول تقنيات الذكاء الاصطناعي. تؤثر هذه التحيزات في كيفية صياغة المعلومات أو إدراجها أو تفسيرها، مما قد يشكل الرأي العام وقرارات السياسة والاعتبارات الأخلاقية المتعلقة بتطوير الذكاء الاصطناعي ونشره.

يؤدي تحيز البلاغة إلى تشكيل التصورات والقرارات المتعلقة بالذكاء الاصطناعي بطرق قد لا تعكس بدقة قدراته أو مخاطره. وقد يؤدي ذلك إلى مفاهيم عامة خاطئة، أو توقعات غير واقعية، أو مخاوف في غير محلها فيما يتعلق بتقنيات الذكاء الاصطناعي. يمكن لمثل هذه التحيزات أن تدفع القرارات التي إما أن تقيد بشكل غير مبرر الاستخدامات المفيدة للذكاء الاصطناعي أو تفشل في معالجة أضراره المحتملة بشكل مناسب.



التأثير

مقالة إخبارية عن تقنية التعرف على الوجوه قد تُظهر تحيزاً حين تؤكد على فوائد التقنية في تعزيز الأمن، مع إغفال المخاوف المتعلقة بالدقة والخصوصية. من خلال التركيز الانتقائي على قصص النجاح والتقليل من أهمية التحديات الأخلاقية، تساهم المقالة في تشكيل تصورات منحازة تتجاهل التعقيدات الحقيقية المرتبطة باستخدام الذكاء الاصطناعي.



المثال

تعزيز الشفافية والتعليم والتمثيل المتنوع في المناقشات، وهذا يعني أن نكون منفتحين بشأن القيود والشكوك المتعلقة بالذكاء الاصطناعي، واثقified الجمهور بشأن مفاهيم الذكاء الاصطناعي وأخلاقياته، وضمان سماع الأصوات المتنوعة. يعد التحقق من الحقائق والمشاركة في حوار بناء أمراً مهماً أيضاً، إلى جانب الالتزام بالإرشادات الأخلاقية للتواصل حول الذكاء الاصطناعي.



التخفيف

تحيز التأكيد

يحدث تحيز التأكيد عندما تميل أنظمة الذكاء الاصطناعي إلى تفضيل المعلومات التي تدعم الفرضيات أو الفئات المسبقة، مع تجاهل أو التقليل من قيمة البيانات التي تتعارض معها. ينشأ هذا التحيز أثناء التدريب أو التقييم، مما يؤدي إلى تشويه في استنتاجات النموذج وقراراته، ويُضعف من دقة الأداء وحيادية النتائج.

يؤدي تحيز التأكيد إلى تعزيز التحيزات أو الافتراضات المسبقة داخل أنظمة الذكاء الاصطناعي، من خلال التركيز فقط على الأدلة التي تدعم هذه الافتراضات. ينتج عن ذلك دورة من التعزيز الذاتي تركز التمييز ضد مجموعات أو أفكار معينة، وتؤدي إلى مخرجات غير عادلة تُسهم في استمرار الصور النمطية وتقويض الموضوعية.



التأثير

نموذج الذكاء الاصطناعي للرعاية الصحية تم تدريبه على التنبؤ بنتائج المرضى بناءً على البيانات الطبية، فإذا كانت مجموعة بيانات التدريب تتكون في الغالب من حالات نجح فيها علاج معين، فقد يطور نموذج الذكاء الاصطناعي تحيز التأكيد من خلال المبالغة في التأكيد على فاعلية هذا العلاج مع تجاهل حالات الفشل. ونتيجة لذلك، قد يتنبأ نموذج الذكاء الاصطناعي بشكل غير دقيق بالنتائج الإيجابية للمرضى الذين يتلقون العلاج حتى في الحالات التي لا يكون ذلك أفضل إجراء متوفر.



المثال

استخدام استراتيجيات مثل تنويع جمع البيانات لتشمل الحالات التي تتحدى المعتقدات الحالية، وضمان اختيار متوازن للميزات لتجنب الاعتماد المفرط على العوامل المؤكدة، واستخدام مقاييس تقييم قوية لتقييم أداء النموذج بموضوعية، وإجراء عمليات تدقيق منتظمة لتحديد التحيزات ومعالجتها.



التخفيف

تحيز التباين

يشير تحيز التباين إلى انحراف يحدث عندما تعتمد أنظمة الذكاء الاصطناعي في تقييماتها أو قراراتها على المقارنات النسبية بين نقاط البيانات، بدلاً من التركيز على الصفات الموضوعية لكل نقطة على حدة. يؤدي هذا التوجه إلى نتائج غير دقيقة، حيث تقيم البيانات بناءً على موقعها ضمن المجموعة وليس على أساس خصائصها الفعلية.

يؤثر تحيز التباين في عدالة أنظمة الذكاء الاصطناعي، خاصة في مجالات مثل التوظيف وتقييم الأداء والإقراض، من خلال اعتمادها على المقارنات بين الحالات بدلاً من التقييم الفردي القائم على معايير موضوعية. يؤدي ذلك إلى نتائج غير عادلة قد تميز ضد أفراد مؤهلين، وتُضخم فروقات لا علاقة لها بالأداء الحقيقي أو الجدارة.



التأثير

في أداة التوظيف المعتمدة على الذكاء الاصطناعي، قد يظهر تحيز التباين في تقييم السيرة الذاتية للمرشح، حيث يقوم النظام بتصنيف المرشحين ليس فقط على أساس مؤهلاتهم ولكن بمقارنتهم مع مجموعة المتقدمين. فقد تتم الاستهانة بالمرشح المؤهل تأهيلاً عالياً إذا تم استعراض سيرته الذاتية في إطار مجموعة قوية بشكل استثنائي من عدد من المتقدمين، مما يؤدي إلى استبعادات غير عادلة تعتمد فقط على الحالة النسبية بدلاً من الجدارة المطلقة.



المثال

تصميم أنظمة الذكاء الاصطناعي لتقييم كل نقطة بيانات بناءً على معايير موضوعية محددة مسبقاً، بشكل مستقل عن العناصر الأخرى لمجموعة البيانات. ويمكن تحقيق ذلك من خلال التعديلات الخوارزمية التي توحد التقييمات وتضمن اتخاذ القرارات على مقاييس مطلقة بدلاً من المقارنات النسبية. بالإضافة إلى ذلك، فإن دمج آليات المراجعة المنتظمة وتعديل المعايير المستخدمة في التقييمات يمكن أن يساعد على الحفاظ على العدالة والموضوعية.



التخفيف

تحيز التجانس الخارجي

يحدث تحيز التجانس الخارجي عندما يميل النظام أو المستخدم إلى النظر إلى أفراد المجموعة الخارجية على أنهم متشابهون بشكل مبالغ فيه، مع تجاهل الفروقات الفردية بينهم، بينما يُنظر إلى أفراد المجموعة الداخلية على أنهم أكثر تنوعاً، مما يؤدي إلى تمثيلات نمطية وغير دقيقة في مخرجات الذكاء الاصطناعي.

يؤدي تحيز التجانس خارج المجموعة إلى تنبؤات وتصنيفات مبسطة بشكل مفرط من قبل أنظمة الذكاء الاصطناعي تجاه الأفراد من خلفيات متنوعة، مما يضعف قدرتها على التمييز بين الفروقات الفردية. يُسفر هذا التحيز عن معاملة غير عادلة، ويحد من فاعلية التدخلات الموجهة، ويُسهّم في ترسيخ الصور النمطية في مجالات حساسة مثل الإعلانات أو التنبؤ القضائي.



نظام ذكاء اصطناعي لتحليل سلوك المستهلكين يصنف جميع العملاء القادمين من قارة معينة كمجموعة واحدة متجانسة، متجاهلاً الاختلافات الكبيرة في العادات والاهتمامات داخل هذه القارة، مما يؤدي إلى حملات تسويقية غير دقيقة وفقدان فرص استهداف فعالة.



الدمج الفعال لمجموعات البيانات المتنوعة والتمثيلية، وإجراء عمليات تدقيق متكررة وشاملة لأنظمة الذكاء الاصطناعي لتحديد ومعالجة أي مظاهر لتحيز التجانس الخارجي، والتأكد من أن مجموعات البيانات والسيناريوهات المتنوعة يتم استخدامها أثناء جميع مراحل الاختبار والتحقق.



تحيز التجميع

يجد تحيز التجميع عندما يتم دمج بيانات من مجموعات مختلفة بطريقة تُغفل الفروقات بينها، مما يؤدي إلى نموذج يخدم المجموعة الأكبر بشكل أفضل ويتجاهل احتياجات المجموعات الأصغر. هذا التحيز ينتج غالباً عن رغبة المطورين في تبسيط البيانات، لكنه يقلل من دقة وعدالة أداء النظام عبر جميع الفئات.

يؤثر تحيز التجميع في عدالة وكفاءة أنظمة الذكاء الاصطناعي من خلال تحسين النماذج لأداء الأغلبية، مع إغفال تمثيل مجموعات الأقليات. يؤدي هذا إلى مخرجات غير دقيقة أو غير منصفة لفئات مهمشة، مما يسهم في استمرار الفوارق القائمة أو تفاقمها، ويقلل من فاعلية الأنظمة في معالجة التنوع البياني بشكل عادل.



التأثير

نظام ذكاء اصطناعي لتحليل زيادات الرواتب قد يظهر تحيزاً عند دمج مهن مختلفة دون مراعاة خصائصها الخاصة، فقد يستنتج النموذج أن الأجر يرتفع طردياً مع سنوات الخبرة، كما هو الحال في مجالات مثل التمويل والتعليم. لكن هذا النمط لا ينطبق على الرياضيين، الذين يتقاضون رواتب عالية في بداية مسيرتهم وتنخفض لاحقاً مع تراجع لياقتهم. بدمج هذه المهن معاً دون تمييز، يصبح النموذج متحيزاً ضد الرياضيين ويعطي نتائج غير دقيقة.



المثال

إنشاء نماذج مختلفة مخصصة لمجموعات معينة أو دمج النماذج باستخدام أساليب التجميع لتحسين الأداء عبر مجموعات سكانية مختلفة. علاوة على ذلك، يستحسن التأكد من أن طريقة جمع البيانات تشمل عينة متنوعة، بالإضافة إلى ذلك، فإن استخدام معايير العدالة لتقييم أداء النموذج عبر المجموعات السكانية المختلفة، وإجراء عمليات تدقيق متكررة، يمكن أن يساعد على تحديد وتصحيح تحيز التجميع.



التخفيف

تحيز التحول المتكرر في المفهوم

يشير إلى الظاهرة التي تحدث فيها التغييرات في الأنماط الأساسية أو العلاقات في البيانات التي يعتمد عليها نموذج الذكاء الاصطناعي للتنبؤات بشكل دوري. يتميز هذا النوع من الانحراف بتكرار حدوث ظروف أو سلوكيات معينة تؤثر في أداء النموذج، عاكساً التقلبات المنهجية التي تحدث عبر الزمن، مثل الاتجاهات الموسمية أو الأحداث الدورية.

يؤثر في القدرة على التنبؤ، في حين أن الطبيعة المنهجية لهذه التغييرات تسمح بمستوى معين من الترقب والتخطيط، فإن الفشل في أخذها في الاعتبار بشكل مناسب يمكن أن يؤدي إلى انخفاضات دورية في أداء النموذج.



التأثير

في نماذج مبيعات التجزئة، قد ترتفع أحجام المبيعات خلال مواسم العطلات، أو فترات الخصومات بسبب التغييرات المتوقعة في سلوك المستهلك. كما تظهر مبيعات الآيس كريم اتجاهات موسمية واضحة، حيث ترتفع النسبة في أشهر الصيف وتنخفض خلال المواسم الباردة. تمثل هذه الأنماط انحرافاً متكرراً للمفهوم مع تغير بيئة النموذج بطريقة دورية قابلة للتنبؤ، مما يؤثر في أهمية العلاقات المستفادة خلال هذه الفترات.



المثال

إضافة آليات في نموذج الذكاء الاصطناعي الذي يمكنه التكيف مع هذه التغييرات الدورية أو توقعها. ويتضمن استخدام بنى النماذج التي تتعامل مع الموسمية بشكل فعال، مثل نماذج تحليل السلاسل الزمنية التي يمكنها التبديل بين النماذج المختلفة المحسنة لفترات زمنية محددة. كما أنه من الضروري اختيار بيانات التدريب بعناية للتأكد من أنها تعكس الطبيعة الدورية لانحراف المفهوم.



التخفيف

تحيز التحول المفاجئ في المفهوم

يشير إلى تغيير مفاجئ وكبير في الأنماط أو العلاقات الكامنة ضمن البيانات التي يعتمد عليها نموذج التعلم الآلي في تنبؤاته. على عكس الانحراف التدريجي، يحدث الانحراف المفاجئ بشكل غير متوقع، مما يؤدي إلى تراجع سريع في أداء النموذج حيث لم تعد الأنماط المتعلمة سابقاً تنطبق على البيئة الجديدة.

يؤثر تحيز التحول المفاجئ في المفهوم بشكل كبير على فاعلية نماذج الذكاء الاصطناعي، حيث يمكن أن تصبح غير دقيقة أو غير صالحة للاستخدام خلال فترة زمنية قصيرة. تؤدي هذه التغيرات إلى تعطيل العمليات واتخاذ قرارات ضعيفة، مما يستدعي تدخلاً سريعاً لإعادة ضبط النموذج وضمان استجابته للسياقات الجديدة والمتغيرة.



التأثير

نماذج التعلم الآلي تأثرت بشكل كبير بجائحة كوفيد-19، التي شكلت مثلاً واضحاً على الانحراف المفاجئ للمفاهيم. على سبيل المثال، واجهت نماذج التنبؤ بتوافر المواد الغذائية صعوبات كبيرة، حيث أصبحت الأنماط التاريخية للتسوق غير فعالة نتيجة التحولات المفاجئة في سلوك المستهلك واضطرابات سلسلة التوريد، مما أدى إلى انخفاض دقة التنبؤات.



المثال

تطبيق أنظمة مراقبة نموذجية قوية قادرة على اكتشاف الانخفاضات الكبيرة في الأداء بسرعة. ويعد تنفيذ استراتيجيات التكيف مع النماذج الديناميكية -مثل إعادة التدريب السريع باستخدام البيانات الجديدة أو التحول إلى نماذج أكثر قدرة على التكيف- أمراً بالغ الأهمية. بالإضافة إلى ذلك، فإن تطوير النماذج التي تتمتع بالمرونة اللازمة للتكيف مع المعلومات الجديدة يمكن أن يساعد على إدارة تأثير هذا الانحراف.



التخفيف

تحيز التخصيص

يحدث تحيز التخصيص عندما تقوم الخوارزمية بتوزيع الموارد أو الفرص أو الاهتمام بشكل غير عادل بين المستخدمين. قد ينتج هذا التحيز عن بيانات تدريب غير متوازنة، أو تصميم نموذجي يفضل فئة معينة، أو انعكاس غير مقصود لتحيزات موجودة في المجتمع، مما يؤدي إلى تفضيل بعض الفئات على حساب فئات أخرى.

يؤدي إلى توزيع غير عادل للموارد أو الفرص أو الاهتمام، مما ينتج عنه حرمان مجموعات معينة بشكل منهجي مع تفضيل مجموعات أخرى. ويمكن أن يكون لذلك آثار مجتمعية كبيرة، مثل عدم المساواة في الوصول إلى الرعاية الصحية أو الخدمات المالية أو الفرص التعليمية، مما يعزز الفوارق القائمة.



التأثير

يتم تدريب نظام الذكاء الاصطناعي للتوظيف على البيانات التاريخية التي تحتوي على تحيزات تُفضّل تركيبة سكانية محددة، وقد تخصص الخوارزمية توصيات أكثر إيجابية للمرشحين من تلك المجموعات، مما يؤدي إلى استمرار التحيزات السابقة.



المثال

استخدام مجموعات بيانات ممثلة لجميع الفئات بشكل عادل، بحيث يمكن استخدام تقنيات أخذ العينات الطباقية لضمان توزيع البيانات بشكل متساو أثناء تدريب النموذج. تساعد النماذج المناهضة للتحيز على تصحيح التحيزات بشكل كبير. كما أن إجراء عمليات تدقيق منتظمة لقرارات الذكاء الاصطناعي أمر ضروري لاكتشاف ومعالجة أي تفاوتات. بالإضافة إلى ذلك، يمكن تعديل مخرجات الذكاء الاصطناعي بعد التدريب، ويسهم الإشراف البشري في القرارات المهمة في ضمان التوزيع العادل.



التخفيف

تحيز التخوف

يشير إلى نوع من التحيز الذي يمكن أن يحدث عندما تعكس أنظمة الذكاء الاصطناعي، أو البيانات المستخدمة لتدريب هذه الأنظمة، بشكل ضمني أو صريح مخاوف الإنسان، أو أحكام مسبقة، أو افتراضات سلبية. يمكن أن يؤدي ذلك إلى نماذج ذكاء اصطناعي شديدة الحذر، أو تمييزية، أو تلك التي تعزز الصور النمطية وعدم المساواة الاجتماعية.

يؤدي ذلك إلى ظهور أنظمة ذكاء اصطناعي مفرطة في المحافظة، أو تمييزية، أو تعمل على استمرار الصور النمطية الضارة. وهذا يمكن أن يقوض فائدة الأنظمة من خلال منعها من اتخاذ قرارات متوازنة، مما قد يؤدي إلى الإضرار بالمجموعات الخاضعة لهذه الافتراضات السلبية. علاوة على ذلك، يمكن للتحيز في التخوف أن يؤدي إلى تراجع ثقة الجمهور في تقنيات الذكاء الاصطناعي، وخاصة إذا كان يُنظر إليها على أنها تدمج أو تتصرف بناءً على مخاوف أو تحيزات لا أساس لها.



في نظام تنبؤي للشرطة يتم تدريبه على بيانات الجريمة السابقة حيث إن البيانات تعكس ممارسات متحيزة من قبل الشرطة، مثل ارتفاع معدلات الاعتقال في مجتمعات معينة بسبب المبالغة في ضبط الأمن، فقد يتعلم نظام الذكاء الاصطناعي كيفية استهداف هذه المجتمعات بشكل غير متناسب، بافتراضه أن هذه المجتمعات ذات معدلات جريمة مرتفعة بسبب وجود تحيز بالفعل في جمع البيانات، وغالباً لا يكون ذلك بسبب معدلات الجريمة الحقيقية.



تدريب أنظمة الذكاء الاصطناعي باستخدام مجموعات بيانات متنوعة وتمثيلية بحيث تشمل مجموعات البيانات هذه مجموعة واسعة من التركيبة السكانية والخبرات ووجهات النظر لمنع النظام من تبني مخاوف أو تحيزات شبيهة بالبشر. إن استخدام التقنيات لتحديد التحيزات والحد منها بشكل فعال أثناء عملية تطوير الذكاء الاصطناعي يضمن أيضاً أن مخرجات النظام لا تؤدي إلى استمرار الصور النمطية السلبية أو عدم المساواة.



تحيز التذكر

يحدث عندما يكون الأفراد أكثر ميلاً إلى تذكر أحداث أو تجارب معينة عن غيرها، وغالباً ما يكون ذلك بسبب عوامل مثل الانخراط العاطفي أو الحداثة أو التكرار. يمكن أن يؤدي هذا إلى جمع وتحليل البيانات بشكل متحيز، مما يؤثر في دقة نماذج الذكاء الاصطناعي المدربة على تلك البيانات.

يؤدي تحيز التذكر إلى تطوير نماذج ذكاء اصطناعي غير دقيقة أو غير موثوقة نتيجة الاعتماد على بيانات متحيزة أو غير متوازنة. تتسبب هذه النماذج في اتخاذ قرارات منحازة قد تؤدي إلى نتائج سلبية على الأفراد أو المؤسسات، مما يُضعف الثقة العامة في تقنيات الذكاء الاصطناعي ويحد من قبولها في البيئات الحساسة.



عندما تكون هناك دراسة تبحث في العلاقة بين التدخين وسرطان الرئة. قد يطلب الباحثون من المشاركين تذكر تاريخ التدخين لديهم. قد يكون الأفراد الذين تم تشخيص إصابتهم بسرطان الرئة أكثر ميلاً إلى تذكر عادات التدخين والإبلاغ عنها بالتفصيل، في حين قد يكون أولئك الذين لا يعانون من سرطان الرئة أقل ميلاً إلى تذكر تاريخ التدخين بدقة. قد يؤدي هذا الاختلاف في التذكر إلى ارتباط متحيز بين التدخين وسرطان الرئة، مما قد يقلل من تقدير المخاطر الحقيقية.



استخدام مجموعة متنوعة من طرق جمع البيانات للحد من تأثير تحيز التذكر، وطلب تصنيف البيانات من عدة أشخاص لتقليل آثار تحيز التذكر الفردي. كما أن من المهم استخدام مقاييس التقييم التي لا تتأثر بتحيز التذكر.



تحيز الترابط الوهمي

يحدث عندما يحدد نظام الذكاء الاصطناعي، تحت تأثير أنماط بيانات التدريب أو تحيزات مطوريه، علاقة غير صحيحة بين متغيرين غير مرتبطين. يعكس هذا التحيز الخطأ المعرفي البشري في إدراك العلاقات أو الأنماط غير الموجودة في الواقع.

يؤدي هذا التحيز إلى اتخاذ قرارات وتنبؤات خاطئة في مختلف مجالات استخدام الذكاء الاصطناعي. وعند بناء الاستنتاجات على علاقات غير حقيقية، قد تنتج النماذج مخرجات لا تفتقر فقط إلى الدقة، بل قد تكون أيضاً مضللة أو ضارة، مما يسهم في تعزيز الصور النمطية أو نشر افتراضات غير مدعومة، وهو ما قد يؤثر سلباً في القرارات والإجراءات والسياسات المعتمدة على هذه النتائج.



التأثير

عندما يتم تدريب نظام ذكاء اصطناعي على مراجعة طلبات التوظيف لمجموعة بيانات يتكرر فيها ذكر الرياضات الجماعية من قبل المرشحين الناجحين، فقد يطور الذكاء الاصطناعي انحيازاً وهمياً مفترضاً بشكل خاطئ أن ذكر المشاركة في الرياضات الجماعية هو مؤشر للأداء الوظيفي. ونتيجة لذلك، قد يفضل النظام بشكل غير عادل المتقدمين الذين يدرجون الرياضات الجماعية في سيرتهم الذاتية، ويهمل المرشحين الآخرين الذين قد يكونون متساوين في المؤهلات أو أكثر ولكنهم لا يذكرون مثل هذه الأنشطة.



المثال

تحليل البيانات ومعالجتها مسبقاً لتحديد الارتباطات الزائفة واستبعادها قبل أن تؤثر في النموذج. يمكن أن يساعد استخدام الأساليب الإحصائية لاختبار أهمية وسببية العلاقات في البيانات على ضمان أن الارتباطات الصحيحة فقط هي التي توجه عملية تعلم النموذج.



التخفيف

تحيز التساهل

يحدث عندما يعطي نظام الذكاء الاصطناعي تقييمات أعلى مما يستحقه الواقع، ويحدث هذا عادةً لأن البيانات التي تدرب عليها النظام أو الخوارزميات التي يستخدمها تميل إلى تفضيل النتائج الإيجابية. نتيجة لذلك، تظهر النتائج بمستوى أفضل مما هو عليه فعلياً، مما يؤدي إلى قرارات غير دقيقة.

يؤثر تحيز التساهل في دقة التقييمات من خلال الميل إلى إعطاء درجات أعلى مما يستحقه الأداء الفعلي، مما يؤدي إلى طمس الفروقات بين المساهمات المتميزة والعادية. يُقوّض هذا التحيز نزاهة عمليات اتخاذ القرار، وقد يُفضي إلى تخصيص غير عادل للموارد أو الترقيات، على حساب الأفراد الأكثر كفاءة واستحقاقاً.



نظام الذكاء الاصطناعي المستخدم لتصنيف مقالات الطلاب يظهر تحيزاً للتساهل إذا كان يمنح باستمرار درجات أعلى من تلك التي تستحقها المقالات بناءً على المعايير الموضوعية، وقد ينبع هذا من التدريب على مجموعة بيانات ذات درجات عالية للغاية أو خوارزمية تفضل بعض المؤشرات الإيجابية بشكل غير متناسب، مما يجعل الطلاب يحصلون على درجات مبالغ فيها لا تعكس بدقة مهاراتهم في الكتابة، مما يؤثر في تعلمهم وتطورهم.



تطوير وتدريب أنظمة الذكاء الاصطناعي للتأكد من أنها تعكس بدقة المعايير التي من المفترض تقييمها. ويمكن أن يشمل ذلك تنوع مجموعات بيانات التدريب لتشمل مجموعة من الأداءات، وتنفيذ عمليات التحقق الخوارزمية التي تعيد معايرة النقاط لمنع الانحراف نحو التساهل، ودمج آليات للمراجعة المنتظمة والتعديل بناءً على النتائج الموضوعية.



تحيز التسلسل الزمني

يشير إلى الحالة التي تتأثر فيها التنبؤات أو القرارات التي تتخذها أنظمة الذكاء الاصطناعي بتوقيت أو تسلسل الأحداث في البيانات، بدلاً من الاعتماد على الأهمية الجوهرية أو الدلالة للمعلومات. يمكن أن ينشأ هذا التحيز عندما تعكس البيانات التاريخية المستخدمة لتدريب نموذج الذكاء الاصطناعي أنماطاً زمنية معينة أو اتجاهات قد لا تمثل بدقة الحالة الحالية أو المستقبلية للأمور.

يؤثر تحيز التسلسل الزمني في دقة نماذج الذكاء الاصطناعي من خلال دفعها لإعطاء أولوية مفرطة للأنماط التاريخية، حتى في الحالات التي لا تعكس فيها تلك الاتجاهات الواقع الحالي. يؤدي هذا التحيز إلى قرارات غير ملائمة في البيئات سريعة التغير، ويقلل من قدرة النموذج على التنبؤ الدقيق بمواقف لا ترتبط بها النتائج المستقبلية بالماضي القريب.



التأثير

نظام تنبؤ بسوق الأوراق المالية مدعوم بالذكاء الاصطناعي تم تدريبه على بيانات أسعار الأسهم السابقة. إذا كانت مجموعة بيانات التدريب تتكون في الغالب من فترات النمو الاقتصادي، فقد يتعلم نموذج الذكاء الاصطناعي التنبؤ بأسعار الأسهم المستقبلية بناءً على تلك الأنماط السابقة. ولكن إذا تغيرت الظروف الاقتصادية بشكل كبير بعد نشر النموذج؛ على سبيل المثال، إذا كان هناك ركود مفاجئ، فقد يواجه نظام الذكاء الاصطناعي صعوبة في التنبؤ بدقة أسعار الأسهم لأنه متحيز نحو الأنماط السابقة التي لم تعد تنطبق.



المثال

استخدام تقنيات مثل المعالجة المسبقة للبيانات لمراعاة التحولات أو الاتجاهات الزمنية، مع دمج المعلومات الحديثة ضمن مجموعة بيانات التدريب، وإعادة تدريب النموذج بشكل منتظم لضمان قدرته على التكيف مع الظروف المتغيرة والحفاظ على دقته وكفاءته بمرور الوقت.



التخفيف

تحيز التسمية

يشير إلى عدم الدقة في البيانات الموسومة المستخدمة لتدريب نماذج التعلم الآلي، والناجمة من التصورات الذاتية أو التحيزات الكامنة لدى المسمين (Annotators). يحدث هذا التحيز أثناء عملية تسمية البيانات، حيث يصنف الموصِّفون البشريون البيانات بناءً على فهمهم، أو تفسيراتهم، أو تحيزاتهم. نتيجة لذلك، قد لا تمثل بيانات التدريب الظاهرة التي يفترض أن تنمذجها بشكل موضوعي، مما يؤدي إلى أنظمة ذكاء اصطناعي تعكس هذه التحيزات الذاتية.

يؤثر تحيز التسمية بشكل مباشر في عملية تعلم نماذج الذكاء الاصطناعي، من خلال إدخال تصنيفات ناتجة عن تصورات أو تحيزات شخصية لدى المسمين. يؤدي هذا إلى نتائج منحرفة لا تعكس الواقع بدقة، ويقوض عدالة الأنظمة الذكية، مما يُضعف موثوقيتها ويقلل من ثقة المستخدمين في مخرجاتها، خاصة في التطبيقات الحساسة.



التأثير

عند تطوير أنظمة الذكاء الاصطناعي لتصنيف رسائل البريد الإلكتروني العشوائية، قد يؤدي اختلاف معايير التصنيف بين القائمين على إعداد البيانات إلى تحيز في النظام. فعلى سبيل المثال، قد يصنف بعض المصنفين جميع الرسائل الترويجية كبريد عشوائي، في حين يقتصر آخرون على تصنيف رسائل التصيد الاحتيالي أو الرسائل الضارة فقط. ينتج عن هذا التباين تحيز يؤثر سلباً في دقة مرشحات البريد العشوائي، ويؤدي إلى تصنيف غير موثوق للرسائل.



المثال

وضع معايير واضحة وموضوعية لتصنيف البيانات، والتي يمكن أن تساعد على تقليل موضوعية التعليقات التوضيحية وضمان الاتساق بينها. بالإضافة إلى ذلك، يمكن أن يساعد إنشاء فرق من المسمين من خلفيات مختلفة على تقليل التحيزات الفردية، مما يؤدي إلى مجموعة بيانات أكثر توازناً. علاوة على ذلك، فإن استخدام تقنيات مراقبة الجودة، مثل قيام عدد من المسمين بتسمية نفس البيانات والتوفيق بين التباينات بالإجماع أو مراجعة الخبراء، يمكن أن يساعد على تحديد وتصحيح التسميات المتحيزة.



التخفيف

تحيز التصادم

يحدث تحيز التصادم عندما يتم تدريب نموذج الذكاء الاصطناعي على مجموعة بيانات يظهر فيها ارتباط زائف بين متغيرين ليس بسبب علاقة سببية بينهما، بل نتيجة اعتمادهما المشترك على متغير ثالث يُعرف بالمصادم. ويظهرها التحيز عندما يتم تضمين هذا المتغير في التحليل أو التعديل عليه، مما يؤدي إلى تشويه العلاقة الحقيقية بين المتغيرات في النموذج.

يؤدي تحيز التصادم إلى استنتاجات مضللة حول العلاقة بين المتغيرات، مما قد يتسبب في قيام نماذج الذكاء الاصطناعي باتخاذ قرارات بناءً على علاقات ارتباطات زائفة بدلاً من العلاقات السببية الحقيقية. قد يؤدي ذلك إلى تشويه الدقة التنبؤية ويؤدي إلى اتخاذ إجراءات غير مناسبة.



التأثير

في مجموعة بيانات تتنبأ بدرجات اختبارات الطلاب، يرتبط كل من "الوقت الذي يتم قضاؤه في الدراسة" و"ساعات النوم" بأداء الاختبار. ولكن بالإضافة إلى ذلك، فإن الطلاب يتأثرون أيضاً بمتغير خفي: "الاهتمام بالمادة". ويعمل هذا المتغير الخفي بمثابة مُصادم، مما يؤدي إلى التحيز عن طريق إنشاء علاقة مصطنعة بين "الوقت الذي تقضيه في الدراسة" و"ساعات النوم". وإذا لم يأخذ النموذج هذا التحيز في الاعتبار، فقد يفترض افتراضاً خاطئاً بوجود علاقة مباشرة بين الدراسة والنوم، مما يؤدي إلى تنبؤات غير دقيقة.



المثال

تحديد عوامل التصادم المحتملة وفهم تأثيرها في المتغيرات محل الاهتمام، مع الاستفادة من تقنيات مثل أساليب الاستدلال السببي، وتكامل معرفة المجال، وتحليلات الحساسية لمعالجة تحيز التصادم وتعزيز دقة النماذج وموثوقية نتائجها.



التخفيف

تحيز التصنيف الخاطئ

يحدث عندما يقوم نظام الذكاء الاصطناعي بتصنيف نقاط البيانات بشكل غير صحيح مما يؤدي إلى أخطاء في توقع النتائج. هذا النوع من التحيز شائع في التعلم الموجه حيث يتم تصنيف مجموعة البيانات التدريبية. وإذا تم تصنيف بعض البيانات بشكل غير صحيح، فإن ذلك يؤثر بشكل سلبي في مخرجات نظام الذكاء الاصطناعي.

يؤدي تحيز التصنيف الخاطئ إلى أخطاء في كيفية تصنيف نقاط البيانات بواسطة أنظمة الذكاء الاصطناعي، مما يؤثر بشكل مباشر في دقة تنبؤات المخرجات وربما يؤدي إلى قرارات غير عادلة أو غير صحيحة تؤثر في مجموعات معينة أكثر من غيرها.



التأثير

في نظام التعرف على الوجوه، إذا كانت بيانات التدريب لهذا النظام تتضمن عدداً كبيراً من الوجوه لمجموعة عرقية محددة، فقد يصبح الذكاء الاصطناعي أفضل في التعرف على الأفراد من هذه المجموعة بدلاً من الآخرين. ونتيجة لذلك، سيظهر معدل أعلى من تصنيف إيجابي خاطئ أو تصنيف سلبي خاطئ للأفراد الذين لا يكونون ضمن مجموعة الأغلبية في بيانات التدريب (يصبح معدل التشخيص الخاطئ للأقليات أعلى).



المثال

استخدام تقنيات لتحديد وتصحيح مصادر التحيز المحتملة، ويمكن أن يشمل ذلك زيادة البيانات لتحقيق التوازن بين الفئات الممثلة تمثيلاً ناقصاً، والتحقق المتبادل للتحقق من أداء النموذج المتسق عبر مجموعات فرعية مختلفة من البيانات، والتدقيق المنتظم لتنبؤات النموذج بمقارنتها بالبيانات المرجعية. بالإضافة إلى ذلك، دمج حلقات التغذية الراجعة حيث يمكن الإبلاغ عن التصنيفات الخاطئة واستخدامها لتحديث النموذج يمكن أن يساعد على تحسين الدقة وتقليل التحيز بمرور الوقت.



التخفيف

تحيز التطوع

يحدث تحيز التطوع عندما يتم تدريب النموذج على مجموعة بيانات مأخوذة من مشاركين اختاروا المشاركة بأنفسهم أو من نقاط بيانات غير عشوائية. يحدث ذلك عادةً في الاستطلاعات، أو الدراسات، أو أي شكل من أشكال المساهمة الطوعية بالبيانات، حيث يختار الأفراد المشاركة بناءً على اهتماماتهم أو توفرهم أو دوافعهم الشخصية. ونتيجة لذلك، قد تمثل البيانات آراء أو سلوكيات أو خصائص شريحة معينة من السكان ممن لديهم ميل أو قدرة أكبر على المشاركة، بدلاً من تقديم رؤية شاملة وعادلة عن المجتمع ككل.

يؤثر تحيز التطوع في قابلية تعميم وعدالة تنبؤات النموذج وقد تحقق نماذج الذكاء الاصطناعي المدربة على البيانات المتأثرة بتحيز التطوع أداءً جيداً لمجموعة فرعية محددة ممثلة تمثيلاً زائداً في مجموعة البيانات ولكنها تفشل في التنبؤ بدقة أو فهم سلوكيات السكان على نطاق أوسع. ويمكن أن يؤدي ذلك إلى نتائج وقرارات منحرفة لا تأخذ في الاعتبار تنوع الخبرات ووجهات النظر، مما قد يؤدي إلى تهميش الفئات الممثلة تمثيلاً ناقصاً وتعزيز الفوارق.



التأثير

عندما يتم تطوير نظام ذكاء اصطناعي يهدف إلى فهم تفضيلات المستهلك بناءً على استجابات الاستطلاع عبر الإنترنت، إذا تم إكمال الاستطلاع فقط من قبل مجموعة سكانية معينة من المستهلكين الذين هم أكثر خبرة في التقنية أو لديهم وقت فراغ أكبر، فمن المرجح أن يطور نموذج الذكاء الاصطناعي الناتج فهماً متحيزاً لتفضيلات المستهلك. وبالتالي، قد يوصي النموذج باستراتيجيات أو منتجات تلبي تفضيلات المشاركين في الاستطلاع ولكنها تفشل في جذب قاعدة عملاء أوسع، بما في ذلك أولئك الذين لم يشاركوا في الاستطلاع.



المثال

إضافة البيانات من الموارد الإضافية التي لا تعتمد على مشاركة المتطوعين في سد الفجوات وموازنة مجموعة البيانات. قد يتضمن ذلك استخدام السجلات العامة أو بيانات المراقبة أو قواعد بيانات من طرف ثالث. علاوة على ذلك، فإن استخدام التقنيات الإحصائية مثل الترجيح والتقسيم الطبقي لتصحيح مجموعة البيانات للتمثيل الزائد أو الناقص لمجموعات محددة قد يساعد على التخفيف من تأثير تحيز التطوع.



التخفيف

تحيز التغذية الراجعة

يحدث عندما تتم إعادة تدوير مخرجات نظام الذكاء الاصطناعي كمدخلات، مما يؤدي إلى إنشاء دورة يمكن أن تؤدي إلى استمرار التحيزات الحالية وحتى تضخيمها. تنشأ هذه الظاهرة لأن قرارات النظام تؤثر في البيانات اللاحقة التي يتلقاها ويعالجها، مما يؤدي إلى حلقة ذاتية التعزيز حيث يتم تعزيز التحيزات الأولية بشكل مستمر بدلاً من تصحيحها مع مرور الوقت.

يؤدي تحيز التغذية الراجعة إلى تفاقم التحيزات في أنظمة الذكاء الاصطناعي من خلال إعادة استخدام مخرجات النظام كمدخلات مستقبلية، مما يخلق حلقة مغلقة من التعزيز الذاتي. ينتج عن هذا التحيز تراكم تدريجي للانحرافات، حيث تصبح القرارات أقل توازناً بمرور الوقت، ويُضعف ذلك قدرة النظام على التصحيح الذاتي أو التحسن مع التجربة.



التأثير

في أنظمة الشرطة التنبؤية التي تستخدم بيانات الجريمة التاريخية المتحيزة للتنبؤ بمناطق الجريمة المتوقعة في المستقبل، ستتم زيادة التشديد في مراقبة الأحياء بناءً على البيانات الأولية، مما يؤدي ذلك إلى مزيد من الاعتقالات في تلك المنطقة بسبب زيادة تواجد الشرطة، ثم تتم تغذية هذه البيانات مرة أخرى في النظام، مما يعزز التحيز المبني على أن المنطقة ذات احتمالية كبيرة لحدوث الجرائم المستقبلية، ويؤدي ذلك إلى مزيد من نشر الشرطة في تلك المنطقة، وإعادة تكرار وترسيخ هذا المفهوم.



المثال

التحول نحو أنظمة الذكاء الاصطناعي الديناميكية والتكيفية والتي يمكنها التعرف على التغيرات في بيانات الإدخال الخاصة بها والتكيف معها بمرور الوقت، وينطوي ذلك على تطوير آليات لتحديد وكسر حلقات التغذية الراجعة، مثل إعادة التدريب الدوري للنماذج باستخدام مجموعات بيانات متنوعة ومحدثة، وتنفيذ خوارزميات مدركة للإنصاف وقادرة على تعديل معايير اتخاذ القرار لمنع تعزيز التحيزات.



التخفيف

تحيز التغطية

يجدث عندما لا تمثل مجموعة البيانات المستخدمة لتدريب نموذج الذكاء الاصطناعي بشكل كافٍ تنوع السكان أو الظواهر قيد الدراسة. يمكن أن يؤدي ذلك إلى تنبؤات أو قرارات متحيزة تؤثر بشكل غير متناسب في المجموعات غير الممثلة بشكل كافٍ، بمعنى آخر، لم يتم اختيار البيانات بطريقة تمثيلية.

يؤدي تحيز التغطية إلى أداء ضعيف للأنظمة الذكاء الاصطناعي عند التعامل مع أفراد أو سيناريوهات لم تُمَثَّل بشكل كافٍ في بيانات التدريب، مما يُفاقم أوجه عدم المساواة. وتبرز خطورته في القطاعات الحساسة كالرعاية الصحية والمالية والعدالة الجنائية، حيث يمكن أن يُسهم هذا التحيز في تكرار التمييز المنهجي ضد الفئات المهمشة وتقويض العدالة.



التأثير

نموذج للتنبؤ بالمبيعات المستقبلية لمنتج جديد يتم تدريبه باستخدام بيانات استطلاعات هاتفية موجهة فقط للمستهلكين الذين اشتروا المنتج نفسه، دون شمول مستهلكي المنتجات المنافسة. يؤدي هذا القصور في جمع البيانات إلى عدم تمثيل شريحة مهمة من السوق، مما يحد من قدرة النموذج على التنبؤ الدقيق بسلوك السوق الكامل.



المثال

تطبيق استراتيجيات متعمدة عبر مراحل جمع البيانات ومعالجتها وتطوير النماذج. ويشمل ذلك تنوع مصادر البيانات لتشمل تمثيلاً أوسع للسكان المستهدفين، واستخدام تقنيات أخذ العينات الطباقية لتشمل جميع المجموعات الفرعية ذوات الصلة بشكل متناسب، والتقييم المستمر وتحديث أداء النموذج عبر قطاعات متنوعة، وتعزيز تطوير النموذج الشامل من خلال إشراك مجموعة متنوعة من أصحاب المصلحة.



التخفيف

تحيز التقرير

يشير إلى الانحراف الناتج عن طريقة عرض أو تقديم نتائج النماذج أو مخرجات الأنظمة بطريقة غير مكتملة أو انتقائية، مما يؤدي إلى نقل صورة مضللة عن أداء النظام أو دقته. ينشأ هذا التحيز خلال مراحل تطوير النموذج أو أثناء عرض نتائجه، وقد يؤثر في تفسير فاعلية الذكاء الاصطناعي واستخدامه، مما يضعف موثوقية القرارات المبنية عليه.

يؤثر تحيز التقرير في دقة وموضوعية البحوث العلمية من خلال عرض نتائج مختارة أو إغفال بيانات غير مرغوب فيها، مما يؤدي إلى تشويه الفهم الحقيقي للظواهر المدروسة. يُسهم هذا التحيز في اتخاذ قرارات سياسية أو ممارسات عامة غير مستندة إلى أدلة متوازنة، ويقوض مصداقية البحث العلمي ويهدد مسار التقدم المعرفي.



التأثير

نموذج ذكاء اصطناعي للتنبؤ بأسباب حوادث السيارات قد يتأثر بتحيز التقرير عندما يتم تدريبه على بيانات مستمدة من تقارير الشرطة، التي تميل إلى توثيق الحوادث الخطيرة بشكل أكبر من الحوادث البسيطة التي غالباً لا يتم الإبلاغ عنها. نتيجة لذلك، يتعلم النموذج أن الحوادث الخطيرة هي النوع الأكثر شيوعاً، مما قد يؤدي إلى تركيزه في التنبؤات على منع الحوادث الجسيمة، مع إهمال الحوادث الأقل خطورة.



المثال

وضع إجراءات دقيقة لجمع وتحليل البيانات كتضمنين جميع النتائج، بما في ذلك النتائج السلبية أو غير الملائمة، وتوضيح الآثار الجانبية المحتملة بشفافية. كما يمكن تطبيق تصميم الدراسة بشكل منعزل لضمان عدم تأثير توقعات المشاركين أو الباحثين في نتائج الدراسة. بهذه الطرق، يمكن تقليل التحيزات في التقارير وضمان النشر الموضوعي والدقيق للبيانات العلمية.



التخفيف

تحيز التقييم

يحدث تحيز التقييم عندما لا تعكس المعايير أو مجموعات البيانات أو المقاييس المستخدمة لتقييم أداء نموذج الذكاء الاصطناعي تنوع المستخدمين أو السيناريوهات الواقعية. يؤدي ذلك إلى إعطاء نتائج غير دقيقة عن كفاءة النموذج، حيث يبدو أنه يعمل بشكل جيد، بينما يفشل في الأداء بنفس الجودة عند تطبيقه على فئات أو ظروف مختلفة.

يؤدي إلى نشر أنظمة ذكاء اصطناعي غير عادلة، وهذا يعني أنه على الرغم من أن النموذج قد يُظهر دقة عالية في الاختبارات الخاضعة للرقابة، إلا أنه قد يفشل في تحقيق نتائج مماثلة في تطبيقات العالم الحقيقي المتنوعة. يمكن أن يؤدي هذا التناقض إلى تقويض ثقة المستخدم، وربما التسبب في ضرر إذا تم استخدام الذكاء الاصطناعي في مجالات حساسة أو حرجية، مثل الرعاية الصحية أو الخدمات المالية.



التأثير

في نظام الذكاء الاصطناعي للتنبؤ بالجدارة الائتمانية، إذا كان تقييم النموذج يعتمد بشكل غير متناسب على بيانات من سكان المناطق المدنية، دون النظر إلى المناطق القروية، فقد يكون النموذج متحيزاً نحو السلوكيات الائتمانية الشائعة في المناطق المدنية، وبالتالي عند تطبيقه على الأفراد من المناطق القروية، قد يقوم النموذج بتقييم جدارتهم الائتمانية بشكل غير دقيق، إما بحرمان الأفراد المستحقين من الائتمان أو تقديم الائتمان بشروط غير ملائمة، بناءً على تقييم غير كامل لأدائه عبر جميع المجموعات السكانية ذوات الصلة.



المثال

استخدام مجموعة متنوعة من مجموعات بيانات التقييم التي تمثل بدقة النطاق الكامل للسيناريوهات والمستخدمين الذين يهدف نظام الذكاء الاصطناعي إلى خدمتهم، وذلك باستخدام مجموعة واسعة من مقاييس الأداء التي تلتقط جوانب مختلفة من سلوك النموذج، وتوظيف تقنيات التحقق من الصحة التي تختبر أداء النموذج في ظل مجموعة متنوعة من الظروف.



التخفيف

تحيز التوفر

يشير تحيز التوفر إلى ميل أنظمة الذكاء الاصطناعي إلى اعتماد قراراتها أو تنبؤاتها على المعلومات التي يسهل الوصول إليها، بدلاً من الاعتماد على بيانات شاملة وغير متحيزة. يؤدي هذا التحيز إلى تضخيم أهمية أحداث أو أنماط معينة، مما ينعكس على دقة النتائج ويُنتج مخرجات غير متوازنة.

يؤدي إلى اتخاذ قرارات بناءً على شريحة ضيقة من المعلومات، مما ينتج عنه تنبؤات غير شاملة ولا تعكس السياق كاملاً. ويمكن أن يؤدي ذلك إلى اتخاذ قرارات تتأثر بشكل مفرط بالحالات الشاذة الحديثة، ومن المحتمل أن تتجاهل أنماطاً مهمة ولكنها أقل وضوحاً.



التأثير

في نموذج الذكاء الاصطناعي لنظام تقديم توصيات التسوق عبر الإنترنت، إذا كان جمع مجموعة البيانات المستخدمة للتدريب يتم من خلال جمع المعلومات حول تفضيلات المستخدم والمنتجات التي قام بشرائها، وكانت عملية جمع البيانات تركز في المقام الأول على عمليات الشراء الأخيرة وتفاعلات المستخدم، فقد تؤدي إلى تحيز التوفر في مجموعة البيانات.



المثال

تنظيم مجموعات البيانات عمداً لتشمل مجموعة واسعة ومتنوعة من تفاعلات المستخدم، ولا تلتقط الأنشطة الحديثة فحسب، بل تتضمن أيضاً البيانات التاريخية التي تعكس الاتجاهات والأنماط الطويلة المدى. يساعد هذا النهج على موازنة تأثير الأحداث الأخيرة في سياق تاريخي أوسع، مما يضمن عدم المبالغة في تأكيد نظام الذكاء الاصطناعي على الاتجاهات الحالية على حساب الأنماط الراسخة.



التخفيف

تحيز الثقة المفرطة

يشير تحيز الثقة المفرطة إلى ميل نظام الذكاء الاصطناعي أو مطوريه إلى المبالغة في تقدير دقة النموذج أو قدراته الفعلية، مما يؤدي إلى توقعات غير واقعية حول أدائه. يمكن أن ينتج عن هذا التحيز قرارات خاطئة أو الاعتماد المفرط على النظام في مواقف تتطلب تدخلاً بشرياً أو تحليلاً أكثر دقة.

يؤدي تحيز الثقة المفرطة في نماذج الذكاء الاصطناعي إلى قرارات قد تحمل عواقب وخيمة، خصوصاً في المجالات الحساسة مثل الرعاية الصحية، حيث قد يؤدي التشخيص غير الدقيق إلى تدخلات غير مناسبة. كما يُضعف هذا التحيز قدرة المستخدمين على التمييز بين الحالات التي تستدعي الثقة بالنموذج وتلك التي تتطلب مراجعة بشرية، مما يهدد فاعلية النظام ومصداقيته.



التأثير

نظام القيادة الذاتية المدعوم بالذكاء الاصطناعي قد يُبالغ في قدراته عندما يتم الترويج له على أنه قادر على التعامل مع جميع حالات الطقس، بما في ذلك العواصف الثلجية الكثيفة والضباب الكثيف. في الواقع، قد يعجز النظام عن إدراك جميع الظروف الطارئة بدقة أو الاستجابة لها بشكل صحيح، مما قد يؤدي إلى وقوع حوادث أو خسائر نتيجة الثقة المفرطة بقدراته.



المثال

إجراء اختبارات وتقييمات دقيقة للأنظمة الذكاء الاصطناعي لضمان الدقة والموثوقية وتأكيد الشفافية بشأن القيود والقدرات التي تتمتع بها أنظمة الذكاء الاصطناعي. بالإضافة إلى المحافظة على الرقابة البشرية والحكم السليم لمنع الإفراط في الاعتماد على الذكاء الاصطناعي، والأخذ في الاعتبار الآثار الأخلاقية عند تطوير أنظمة الذكاء الاصطناعي ونشرها.



التخفيف

تحيز الجنس

يشير إلى ميل أنظمة الذكاء الاصطناعي لإنتاج نتائج تفضل بشكل غير عادل جنساً على آخر. غالباً ما ينشأ هذا التحيز من بيانات تدريب غير ممثلة، أو التحيزات اللاواعية للمطورين، أو الصور النمطية المجتمعية المدمجة في مصادر البيانات التي يتعلم منها الذكاء الاصطناعي. نتيجة لذلك، قد تكرر نماذج الذكاء الاصطناعي أو تضخم التفاوتات الجنسية القائمة، مما يؤثر في تنبؤاتها، أو توصياتها، أو قراراتها.

يؤدي إلى ممارسات تمييزية، مثل التوظيف التفضيلي لجنس واحد، وتوصيات العلاج الطبي المتهيزة، وتعزيز الصور النمطية. يقوض التحيز على أساس الجنس نزاهة ومصداقية تقنيات الذكاء الاصطناعي، مما قد يؤدي إلى استمرار عدم المساواة وتراجع الثقة في تطبيقاته.



في أدوات فحص السيرة الذاتية الآلية، قد يفضل نظام الذكاء الاصطناعي المرشحين من جنس معين بناءً على بيانات التدريب المتهيزة التي تعكس قرارات التوظيف التاريخية. على سبيل المثال: إذا أظهرت البيانات أن أغلبية الأدوار الفنية يكون موظفوها من الذكور، فقد يؤدي الذكاء الاصطناعي بشكل غير عادل إلى خفض مستوى طلبات التوظيف المقدمة من المرشحات لشغل مناصب مماثلة، على الرغم من مؤهلاتهن.



اتباع نهج شامل، يبدأ بتنوع مجموعات بيانات التدريب للتأكد من أنها تعكس بدقة كلا الجنسين دون تحيز. ويجب على المطورين استخدام تقنيات لتحديد التحيزات بين الجنسين والقضاء عليها بشكل فعال في المعالجة المسبقة للبيانات والتدريب النموذجي. ويمكن أن يشمل ذلك المعالجة المحايدة جنسياً وتطبيق خوارزميات تعلم الآلة المراعية للعدالة. يمكن أن يساعد التدقيق المنتظم لأنظمة الذكاء الاصطناعي فيما يتعلق بالنتائج المتهيزة وآليات ردود الفعل على تعديل النماذج لتجنب التحيز بين الجنسين.



تحيز الحداثة

يحدث ذلك عندما تقوم الخوارزمية بمقارنة الأحداث أو نقاط البيانات الأخيرة بشكل غير متناسب مع الأحداث أو نقاط البيانات القديمة، بغض النظر عن أهميتها النسبية أو ملاءمتها. ويمكن أن يتجلى هذا التحيز في أنظمة الذكاء الاصطناعي التي تحلل الاتجاهات، أو تقوم بالتنبؤات، أو تقيم الأداء، مما يؤدي إلى نتائج متحيزة تفضل ما حدث في الماضي القريب، وينشأ من تصميم النموذج أو تقنيات معالجة البيانات التي تعطي الأولوية للمعلومات الحديثة عن غير قصد.

يؤثر تحيز الحداثة في موثوقية أنظمة الذكاء الاصطناعي من خلال تضخيم أهمية الاتجاهات أو البيانات الحديثة على حساب المعلومات طويلة الأمد. يؤدي هذا التحيز إلى توصيات أو توقعات غير متوازنة، لا تتماشى مع الأهداف الاستراتيجية أو الرؤية الشاملة، مما يُضعف دقة القرارات في مجالات مثل التخطيط المالي أو تقييم الأداء المؤسسي.



في نماذج التنبؤ بسوق الأوراق المالية، قد يؤدي تحيز الحداثة إلى تركيز الذكاء الاصطناعي بشكل مفرط على أحدث التقلبات في أسعار الأسهم، وتجاهل الاتجاهات الطويلة المدى والبيانات التاريخية التي قد تقدم رؤية أكثر توازناً لإمكانات السهم، وهذا يمكن أن يدفع المستثمرين إلى اتخاذ قرارات بناءً على التقلبات القصيرة الأجل بدلاً من استراتيجيات الاستثمار السليمة.



تعديل آليات ترجيح النموذج لتقليل التأثير غير المبرر للأحداث الأخيرة أو من خلال دمج التقنيات التي تصمم بشكل واضح الجوانب الزمنية للبيانات لفهم أهمية الاتجاهات مع مرور الوقت. يمكن أن تساعد أيضاً المراجعة المنتظمة وتحديث عوامل نظام الذكاء الاصطناعي وبيانات التدريب على إعادة ضبط التركيز.



تحيز الخوارزمية

يشير إلى التمييز الممنهج وغير العادل الذي يحدث ضمن الأنظمة الخوارزمية. يمكن أن يتجلى هذا التحيز عندما تنتج الخوارزميات نتائج متحيزة بسبب الافتراضات المعيبة في تصميمها، أو بيانات التدريب المتحيزة، أو تفسير البيانات بشكل خاطئ من قبل الخوارزميات نفسها. ويشمل مجموعة من التحيزات، بما في ذلك تلك المتعلقة بالعرق، والجنس، والعمر، والوضع الاقتصادي، ويمكن أن يؤثر بشكل كبير في عمليات صنع القرار.

يؤدي التحيز الخوارزمي إلى التمييز المنهجي عبر مختلف القطاعات، مما يؤثر في وصول الأفراد إلى الخدمات والفرص والمعاملة العادلة، فهو يقوض الموضوعية والعدالة المتوقعة من الأنظمة الآلية، مما يؤثر في عملية صنع القرار في المجالات الحيوية مثل التوظيف والرعاية الصحية والعدالة والتمويل.



التأثير

نظام لتحديد حدود الائتمان لبطاقات الدفع أظهر تحيزاً عندما قدم حدوداً ائتمانية أعلى للرجال مقارنة بالنساء، رغم تشابه ملفاتهم المالية. من المرجح أن الخوارزمية استندت إلى متغيرات متحيزة لصالح الرجال، مما أدى إلى نتائج غير عادلة تجاه النساء. كما أن غياب الشفافية حول طريقة عمل النظام جعل من الصعب تحديد مصدر التحيز بدقة، مما أثار تدقيقاً عاماً ومطالبات بمراجعة الآلية لمنع التمييز بين الجنسين.



المثال

اتباع نهج شامل يبدأ بتنوع مجموعات بيانات التدريب للتأكد من أنها تعكس تنوع العالم الحقيقي بدقة. يجب على المطورين أيضاً استخدام تقنيات الكشف عن التحيز والتخفيف من حدته طوال دورة حياة تطوير الذكاء الاصطناعي، بما في ذلك عمليات التدقيق المنتظمة للأنظمة للذكاء الاصطناعي بحثاً عن النتائج المتحيزة.



التخفيف

تحيز الرغبة الاجتماعية

ينشأ عندما تعتمد النماذج على بيانات يقدمها الأفراد بطريقة تعكس ما يعتقدون أنه مقبول اجتماعياً بدلاً من التعبير عن سلوكهم أو آرائهم الحقيقية. يؤدي هذا التحيز إلى تزويد أنظمة الذكاء الاصطناعي ببيانات غير دقيقة أو متحيزة، مما يؤثر في جودة التدريب ودقة التنبؤات، خاصة في التطبيقات التي تعتمد على استبانات المستخدمين أو تقييمات العملاء أو الدراسات السلوكية.

يؤثر في دقة وموثوقية الرؤى والتنبؤات والتوصيات المستندة إلى الذكاء الاصطناعي. يمكن أن يؤدي تحيز الرغبة الاجتماعية إلى تطوير نماذج لا تعكس بدقة الأنماط أو التفضيلات الحقيقية. وهذا بدوره يمكن أن يؤثر في فاعلية التدخلات المستهدفة، أو الخدمات الشخصية، أو صنع السياسات بناءً على تحليلات الذكاء الاصطناعي، مما قد يؤدي إلى قرارات واستراتيجيات مضللة.



في أداة تقييم الصحة العقلية المدعومة بالذكاء الاصطناعي، قد لا يقوم المستخدمون بالإبلاغ عن أعراض الاكتئاب بسبب الرغبة في الظهور بمظهر أكثر استقراراً نفسياً. ويمكن أن يؤدي هذا النقص في الإبلاغ إلى تحريف فهم نموذج الذكاء الاصطناعي لحالات الصحة العقلية الحقيقية، مما يؤدي إلى تقييمات أو توصيات أقل دقة للرعاية. ويؤثر هذا التحيز في قدرة الأداة على تحديد الأفراد المحتاجين إلى الدعم بشكل فعال، حيث إنها تبني تنبؤاتها ونصائحها على بيانات قد لا تعكس بدقة الظروف الفعلية للمستخدمين.



تطوير الخوارزميات التي يمكنها اكتشاف وتصحيح التناقضات أو الأنماط التي تشير إلى تحيز الرغبة الاجتماعية في البيانات على تحسين دقة المعلومات التي تتم معالجتها. بالإضافة إلى استخدام استراتيجيات لتشجيع الاستجابات الصادقة وتقليل الضغط للتوافق مع الأعراف الاجتماعية. ويمكن أن يشمل ذلك إخفاء الهوية في إجابات الاستطلاع، أو استخدام تقنيات الاستجواب غير المباشرة، أو تطبيق التعديلات الإحصائية لمراعاة التحيزات المحتملة في البيانات المبلغ عنها ذاتياً.



تحيز الضبط غير المتزامن

يحدث عندما تتم مقارنة نتائج مجموعة معالجة أو تجريبية مع مجموعة ضابطة تم جمع بياناتها في فترة زمنية مختلفة. هذا النوع من التحيز يمكن أن يؤثر بشكل كبير في نتائج الدراسات أو التحليلات في مجال الذكاء الاصطناعي، خصوصاً عندما تكون الظروف الزمنية أو السياقية قد تغيرت بين الفترات الزمنية التي تم فيها جمع البيانات للمجموعتين.

يؤثر تحيز الضبط غير المتزامن في دقة نتائج الدراسات أو النماذج التي تعتمد على مقارنة بين مجموعات بيانات تم جمعها في فترات زمنية مختلفة. قد تؤدي التغيرات في العوامل الخارجية مثل السياسات أو التقنية أو الظروف الاقتصادية إلى تشويه النتائج، مما يُقلل من موثوقية التحليلات ويُضعف صلاحية استنتاجات الذكاء الاصطناعي في السياقات المتغيرة.



نظام ذكاء اصطناعي لتشخيص مرض معين تم تدريبه واختباره باستخدام بيانات مرضى عام 2020م، وأظهر دقة عالية في نتائجه. وعند تقييم أدائه من خلال مقارنة تشخيصاته مع بيانات مرضى من عام 2015م، أظهر تحيزاً في النتائج. يعود ذلك إلى التغيرات الكبيرة التي طرأت على أساليب التشخيص والعلاجات والممارسات الصحية بين الفترتين، مما جعل المقارنة بين المجموعتين غير دقيقة وأثر في موثوقية التقييم.



جمع البيانات المتزامنة بالتأكد من أن البيانات للمجموعات الضابطة والتجريبية تُجمع في نفس الفترة الزمنية لضمان تجانس الظروف. بالإضافة إلى استخدام مجموعات ضابطة تاريخية بحذر حيث إن كان لا بد من استخدام مجموعات ضابطة تاريخية، ينبغي إجراء تحليلات لضبط الفروقات الزمنية والسياقية بين الفترات المختلفة.



تحيز العشوائية

يشير إلى حالة يقدم فيها النموذج نتائج مختلفة في كل تنفيذ، حتى عندما تظل المدخلات كما هي. وتنشأ من العشوائية الكامنة في بعض خوارزميات أو عمليات تعلم الآلة، مما يؤدي إلى تباين في نتائج النموذج. ومن المهم تجنب السلوك العشوائي لتعزيز إمكانية التكرار، والحصول على نتائج أكثر موثوقية وثباتاً من النماذج.

يؤثر وجود العشوائية في نماذج الذكاء الاصطناعي بشكل كبير في إمكانية التنبؤ بها وموثوقيتها وإمكانية تكرار نتائجها. في حين أن العمليات العشوائية يمكن أن تساعد النماذج على استكشاف نطاق أوسع من الحلول، إلا أنها يمكن أن تجعل من الصعب أيضاً تكرار النتائج تماماً، مما يعقد جهود تصحيح الأخطاء والاختبار والتحقق.



نظام روبوت محادثة لخدمة العملاء مصمم للاستجابة للاستفسارات المصرفية قد يُظهر تفاوتاً في الردود بسبب الطبيعة العشوائية لعملية التدريب. فعند طرح نفس الاستفسار في جلسات مختلفة، مثل سؤال عن كيفية فتح حساب توفير جديد، قد يقدم النظام مرة رابطة مباشرة إلى صفحة الطلب، ومرة أخرى شرحاً مطولاً عن فوائد الحساب قبل توفير الرابط، رغم تطابق صيغة السؤال. هذا التفاوت يؤثر في تناسق تجربة المستخدم.



إجراء دورات تدريبية متعددة لفهم السلوك العشوائي بشكل أفضل، واستخدام أساليب المجموعة أو متوسط التوقعات لتحقيق الاستقرار في النتائج كما يسمح استخدام بذرة عشوائية ثابتة بتهيئة نموذج متسق ونتائج عبر عمليات التشغيل. ويمكن أن يؤدي تجميع النتائج من عمليات التشغيل المختلفة إلى تقليل التأثيرات العشوائية، مما يوفر تنبؤات أكثر موثوقية.



تحيز العمر

يحدث تحيز العمر عندما تميز الخوارزميات ضد الأفراد الأكبر سناً، نتيجة التمثيل الزائد للفئات العمرية الأصغر في بيانات التدريب. ينشأ هذا التحيز إما من استبعاد غير مقصود للفئات الأكبر سناً، أو من تصميم لا يراعي الفروقات المتعلقة بالعمر، مما يؤدي إلى نتائج غير دقيقة أو غير عادلة.

يؤدي إلى نماذج أقل دقة أو عدلاً عند إجراء تنبؤات حول الأفراد الأكبر سناً كما أن الآثار المترتبة على التحيز على أساس السن في الذكاء الاصطناعي عميقة، وتمتد إلى مجالات أخلاقية واجتماعية ومهنية حيث يمكن أن تؤدي إلى معاملة غير عادلة لكبار السن في مختلف التطبيقات.



التأثير

نظام الذكاء الاصطناعي الذي تم تطويره لفحص المتقدمين على الوظيفة قد يظهر تحيزاً للعمر من خلال إعطاء الأولوية للخصائص أو المؤهلات الأكثر شيوعاً لدى المتقدمين الأصغر سناً، مثل: الإنجازات التعليمية الحديثة أو الإلمام ببعض الاتجاهات التقنية، وقد يؤدي ذلك إلى حصول المرشحين الأكبر سناً على مرتبة أقل بشكل غير عادل، على الرغم من تمتعهم بالخبرة والمهارات ذات الصلة، وذلك لأن النموذج تعلم ربط ملائمة الوظيفة بالسمات الأكثر انتشاراً بين الأفراد الأصغر سناً.



المثال

إعطاء عمليات تطوير الذكاء الاصطناعي الأولوية لإدراج الفئات العمرية المتنوعة في مجموعات بيانات التدريب، مما يضمن أن البيانات تعكس النطاق الكامل للسكان الذين تخدمهم. يتضمن ذلك البحث في بيانات الأفراد الأكبر سناً ودمجها وتصميم خوارزميات حساسة وقادرة على تفسير الاختلافات المرتبطة بالعمر بدقة. يمكن أن يساعد تنفيذ تقنيات العدالة والكشف عن التحيز، مثل مقاييس العدالة وعمليات تدقيق التحيز، على تحديد ومعالجة التحيز على أساس العمر في نماذج الذكاء الاصطناعي.



التخفيف

تحيز العينة

يحدث عندما لا تلتقط البيانات المختارة للتدريب بشكل تمثيلي التنوع الكامل للمجال الذي من المفترض أن تمثله. يمكن أن يقدم هذا الاختيار غير العشوائي أو المتحيز أخطاء منهجية، حيث يستند تعلم النموذج إلى رؤية غير كاملة أو متحيزة للمشكلة.

يؤثر تحيز العينة في قدرة النموذج على التعميم، إذ يقدم تنبؤات دقيقة ضمن بيانات مألوفة لكنه يفشل عند تطبيقه على حالات خارج العينة. يؤدي ذلك إلى قرارات متحيزة أو غير عادلة، خاصة في المجالات الحساسة، ويقوض مصداقية أنظمة الذكاء الاصطناعي ويحد من فعاليتها في تحقيق العدالة وتقديم فوائد واسعة.



التأثير

نموذج للتنبؤ بالمبيعات المستقبلية لمنتج جديد يتم تدريبه باستخدام بيانات استطلاعات هاتفية أجريت مع مستهلكين اشتروا المنتج ومستهلكين اشتروا منتجاً منافساً. بدلاً من اختيار العينة بشكل عشوائي، تم الاعتماد على أول 200 مستهلك استجابوا لرسائل البريد الإلكتروني، وهم على الأرجح أكثر حماساً من بقية المشتريين العاديين. أدى هذا الانحراف في اختيار العينة إلى احتمالية تحيز البيانات، مما أثر في دقة توقعات النموذج وموثوقية نتائجه.



المثال

تجنب أخذ العينات الملائمة عن طريق تحديد المجموعات المستهدفة بدقة وجمع البيانات المتنوعة. واستخدام طرق أخذ العينات العشوائية لضمان دراسة خالية من التحيز. كما توفر العينة العشوائية البسيطة احتمالات اختيار متساوية لجميع الأعضاء، في حين تضمن العينة العشوائية الطباقية التمثيل عبر المجموعات السكانية الرئيسية.



التخفيف

تحيز القياس

يحدث تحيز القياس في أنظمة الذكاء الاصطناعي عندما تتم المبالغة أو التقليل بشكل منهجي من القيم الحقيقية للبيانات المستخدمة في التدريب أو التقييم. وغالباً ما تنشأ أخطاء القياس كمصدر رئيسي للتحيز، نتيجة عدم الدقة في جمع أو تسجيل البيانات، مما يؤدي إلى نماذج تتعلم من معلومات غير دقيقة، ويؤثر سلباً في أداء النظام ومصادقية مخرجاته.

يؤثر تحيز القياس في دقة ومصادقية نماذج الذكاء الاصطناعي من خلال تدريبها على بيانات غير ممثلة للقيم الحقيقية، مما يؤدي إلى نتائج خاطئة أو مضللة. يظهر هذا التأثير بوضوح في المجالات الحساسة كالرعاية الصحية والتمويل، حيث يمكن أن تسفر التقديرات غير الدقيقة عن تشخيصات خاطئة أو قرارات غير عادلة أو خسائر مالية.



نظام ذكاء اصطناعي يعتمد على بيانات أوزان المرضى قد يتعرض لتحيز في القياس إذا تم جمع الأوزان باستخدام ميزان معيب دون ملاحظة الخل. يؤدي ذلك إلى تسجيل بيانات غير دقيقة، مما يدفع النظام إلى تصنيف المرضى في فئات وزن خاطئة، ويؤثر سلباً في دقة التنبؤات والقرارات المستندة إلى هذه البيانات.



دمج البيانات من عدة مصادر وأجهزة للتخفيف من آثار التحيز من مصدر واحد، مما يؤدي إلى تصوير أكثر دقة لظروف العالم الحقيقي. بالإضافة إلى أن استخدام إجراءات فعالة لتنظيف البيانات والمعالجة المسبقة لاكتشاف وإصلاح العيوب أو عدم الاتساق في البيانات قد يساعد على تقليل تحيز القياس. علاوة على ذلك، فإن معايرة أدوات جمع البيانات والتحقق من البيانات التي تم الحصول عليها مقابل المعايير المعترف بها قد تساعد على ضمان أن تكون بيانات المدخلات عالية الجودة ودقيقة.



تحيز الكشف

ينشأ هذا التحيز عندما تعتمد الخوارزمية على بيانات تم جمعها بأساليب كشف غير موحدة، مما يؤدي إلى تعلم أنماط غير دقيقة أو غير متوازنة تؤثر في أداء النموذج ودقته عبر مختلف الفئات المستهدفة. وتؤدي هذه الاختلافات المنهجية أثناء إدخال بيانات التدريب أو التقييم في صعوبات أثناء تحديد أو تشخيص أو التحقق من النتائج.

يؤثر تحيز الكشف في دقة نتائج أنظمة الذكاء الاصطناعي، خاصة في المجالات التي تتطلب قياساً دقيقاً للنتائج مثل الرعاية الصحية أو العدالة الجنائية. قد يؤدي إلى استنتاجات مضللة حول مدى انتشار المشكلات أو فاعلية التدخلات عبر المجموعات المختلفة، مما يقوض موثوقية القرارات المعتمدة على هذه النماذج.



التأثير

في نماذج الذكاء الاصطناعي المستخدمة للتنبؤ بتفشي الأمراض، إذا كانت البيانات التي يتم إدخالها في نموذج الذكاء الاصطناعي تأتي من مناطق ذات مستويات مختلفة عن مدى إمكانية وصولهم إلى الرعاية الصحية، فقد يتم تشخيص الأمراض أو الإبلاغ عنها بشكل متكرر في المناطق التي تتمتع ببنية تحتية أفضل. وبالتالي، قد يتنبأ الذكاء الاصطناعي بشكل خاطئ بمعدل انتشار المرض في هذه المناطق، ولا يعني ذلك بأن المرض أكثر شيوعاً في هذه المناطق، بل يدل على أن اكتشافه في هذه المناطق أكثر احتمالاً.



المثال

التأكد من أن البيانات المستخدمة لتدريب أنظمة الذكاء الاصطناعي يتم جمعها ومعالجتها بشكل موحد عبر جميع المجموعات، دون اختلافات في طرق أو معايير الكشف. وقد يستلزم ذلك توحيد معايير التشخيص، واستخدام أساليب لضبط التحيزات المعروفة في جمع البيانات، وتنويع مصادر البيانات لتمثيل معدلات الكشف المختلفة بدقة.



التخفيف

تحيز المراجع المتباينة

يحدث عند استخدام مجموعات بيانات مرجعية غير متوازنة أو متحيزة عند تدريب أو تقييم نموذج الذكاء الاصطناعي. يحدث هذا النوع من التحيز عندما تكون مجموعة البيانات المستخدمة لتقييم أداء النموذج أو لمقارنة أدائه مع نماذج أخرى غير ممثلة بشكل عادل أو تعكس تفضيلات أو تحيزات معينة.

يؤثر تحيز المراجع المتباينة في دقة التقييمات والتشخيصات التي تعتمد على بيانات مرجعية غير متوازنة أو منحازة، مما يؤدي إلى ارتفاع معدلات الخطأ. يظهر هذا التحيز بشكل خاص عندما تكون الحالات التي يتم تقييمها غير ممثلة جيداً في بيانات الاختبار، مما يضعف أداء النموذج ويقلل من موثوقية المقارنة بين النماذج المختلفة.



التأثير

نظام ذكاء اصطناعي مصمم لتشخيص الأمراض الجلدية، وتم تدريب هذا النظام على مجموعة بيانات تحتوي على نسبة كبيرة من صور الأمراض الجلدية الشائعة في منطقة جغرافية معينة، بحيث يصبح النظام متحيزاً نحو تشخيص الأمراض الشائعة في تلك المنطقة فقط. وعند تقييم أدائه باستخدام مجموعة بيانات مرجعية تحتوي على نفس النمط الجغرافي، سيبدو أداء النظام جيداً. ومع ذلك، إذا تم استخدام مجموعة بيانات مرجعية تحتوي على أمراض جلدية من مناطق جغرافية مختلفة، فقد ينخفض أداء النظام بشكل كبير.



المثال

تنوع البيانات التدريبية، والتحقق من توازنها، وإجراء التقييم عبر مجموعات بيانات متعددة. كما يساهم الكشف المبكر عن التحيز وتصحيحه في تعزيز دقة أداء الأنظمة وضمان شموليتها لمختلف الفئات والمواقف.



التخفيف

تحيز المراجعة

يشير تحيز المراجعة في أنظمة الذكاء الاصطناعي إلى الانحراف الذي يحدث عند تقييم أداء النماذج أو تصنيف نتائجها باستخدام معايير غير عادلة أو غير موضوعية. يؤدي هذا التحيز إلى تشويه التقدير الواقعي لجودة النموذج أو دقته، مما قد ينتج عنه قرارات غير دقيقة في تحسين النماذج أو اعتمادها في التطبيقات العملية.

يؤثر تحيز المراجعة في عدالة التقييمات وموثوقية القرارات داخل المؤسسات، حيث يؤدي استخدام معايير غير موضوعية إلى نتائج غير دقيقة أو غير منصفة. يمتد تأثيره إلى ثقافة المنظمة والامتثال القانوني والسمعة، كما يُضعف الابتكار والتطوير الشخصي ويقوض الثقة في نظم التقييم والمساءلة.



التأثير

في مؤسسة تعتمد على نظام ذكاء اصطناعي لمراجعة أداء الموظفين، قد تُقيّم نماذج العمل بناءً على معايير غير شفافة أو غير متسقة، مثل عدد الرسائل المرسلة بدلاً من جودة المحتوى أو أثره. هذا يؤدي إلى تقييم غير دقيق لمخرجات الموظف، مما يُضعف قرارات الترقية أو التحفيز. يترتب على ذلك إضعاف الثقة في النظام، وتقليل دافعية الابتكار، وزيادة شعور الموظفين بعدم العدالة، مما يؤثر في ثقافة المنظمة ككل.



المثال

استخدام معايير تقييم موحدة وواضحة وتشكيل لجان تقييم متنوعة. بالإضافة إلى تبني إجراءات التعمية في المراجعة وتدريب المراجعين وزيادة الوعي بالتحيزات. كما يمكن إجراء فحوصات دورية وتكرار التقييمات. وكذلك زيادة الشفافية والمساءلة في العمليات التقييمية. باستخدام هذه الاستراتيجيات، يمكن تقليل التحيزات في التقييمات وتعزيز العدالة والموضوعية.



التخفيف

تحيز المراقب

يشير تحيز المراقب في أنظمة الذكاء الاصطناعي إلى ميل الأفراد إلى تفسير أداء أو تصرفات النماذج بناءً على أفكارهم المسبقة أو توقعاتهم الخاصة، والتي قد تتأثر بأهداف أو مصالح معينة. يؤدي هذا التحيز إلى تقييم غير موضوعي للذكاء الاصطناعي، سواء خلال مراحل تطويره أو اختباره أو الترويج له، مما قد يشوه الصورة الحقيقية لقدرات النظام وحدوده.

يؤثر تحيز المراقب في عدالة ودقة وموثوقية أنظمة الذكاء الاصطناعي. ويثير تحيز المراقب القلق بشكل خاص لأنه يمكن أن يؤدي إلى نماذج الذكاء الاصطناعي التي لا تعكس بشكل موضوعي العالم الحقيقي، ولكنها تتشكل من خلال وجهات نظر ذاتية لمطورها أو مستخدميها من البشر.



التأثير

عندما تقوم شركة تقنية كبرى بتطوير نظام مركبات ذاتية القيادة. وتتلقى الشركة تمويلاً كبيراً من شركة نفطية كبيرة تعطي الأولوية للمقاييس التي تفضل كفاءة الوقود والحد من الازدحام المروري، في حين تقلل من أهمية المقاييس المتعلقة بالتأثير البيئي أو استهلاك الطاقة.



المثال

تضمن مجموعة متنوعة من الأفراد في عمليات جمع البيانات لتقليل التحيزات الفردية في نظام الذكاء الاصطناعي. كما أن استخدام الأساليب الآلية لجمع البيانات يمكن أن يقلل من تأثير التحيز البشري، على الرغم من أنه يجب توخي الحذر لضمان عدم استمرار الخوارزميات المستخدمة للتحيزات.



التخفيف

تحيز المركزية

يحدث تحيز المركزية عندما تميل أنظمة الذكاء الاصطناعي، خاصة في تقييم الأداء أو اتخاذ القرار، إلى إصدار أحكام معتدلة تتجنب القيم القصوى، نتيجة الغموض في البيانات أو اتباع نهج خوارزمي محافظ. يعكس هذا التحيز نمطاً مشابهاً في السلوك البشري، حيث يُفضل الخيار "الآمن" عند غياب معايير واضحة للتمييز.

يؤثر تحيز المركزية في دقة تقييم الأداء من خلال الميل إلى تجاهل الفروق بين المستويات العالية والمنخفضة، مما يؤدي إلى ضغط النتائج نحو الوسط. يُضعف هذا التحيز التقدير العادل للأداء الاستثنائي أو الضعيف، ويؤثر سلباً في قرارات الترقيات والمكافآت والتدخلات، ويقلل من فاعلية أنظمة التقييم المؤسسية.



التأثير

نموذج الذكاء الاصطناعي الذي يستخدم لمراجعة مكالمات خدمة العملاء، قد يؤدي التحيز المركزي فيه إلى تصنيف غالبية المكالمات على أنها متوسطة المستوى، بغض النظر عن الجودة الفعلية للخدمة المقدمة، ويمكن أن يحدث هذا إذا تم تدريب النموذج على مجموعة بيانات مع أمثلة غير كافية للمكالمات ذات الجودة العالية والأقل جودة أو إذا كانت الخوارزمية شديدة الحذر في معايير التصنيف الخاصة بها، مما يؤدي إلى تجميع التقييمات حول النقطة المركزية للمقياس.



المثال

تحسين نظام الذكاء الاصطناعي لضمان قدرته على تعيين مجموعة واسعة من النتائج التقييمية بثقة، مما يعكس بدقة التوزيع الحقيقي لمستويات الأداء. ويمكن تحقيق ذلك من خلال تعزيز مجموعة بيانات التدريب بأمثلة واضحة وموزعة جيداً عبر نطاق الأداء ومن خلال تصميم خوارزمية التقييم للتمييز بشكل أكثر جرأة بين مستويات الإنجاز المختلفة.



التخفيف

تحيز النطاق

يشير إلى الظاهرة التي يتأثر فيها أداء نموذج الذكاء الاصطناعي بخصائص معينة للسكان أو الظروف المضمنة في مجموعة بيانات التدريب الخاصة به، حيث يحدث هذا التحيز عندما تمثل مجموعة البيانات نطاقاً محدوداً من الحالة، مما يؤدي إلى أن يحقق الذكاء الاصطناعي أداءً جيداً في الحالات المشابهة، لكنه يواجه صعوبة في الأداء عند التعامل مع حالات خارج هذا النطاق.

يؤثر تحيز النطاق في دقة أنظمة الذكاء الاصطناعي، من خلال الحد من قدرة النماذج على التعامل مع الحالات التي لا تنتمي إلى نطاق البيانات التي تم تدريبها عليها. يؤدي ذلك إلى توصيات غير مناسبة أو خاطئة، ويُسهم في تقديم أداء أقل فاعلية للبيانات من خارج النطاق الممثل.



نظام الذكاء الاصطناعي لتشخيص سرطان الجلد تم تدريبه في الغالب على صور إصابات لأفراد ذوي بشرة فاتحة، بحيث يصبح أقل دقة في تشخيص الحالات لدى الأفراد ذوي البشرة الداكنة، ويمكن أن تؤدي مجموعة التحيز هذه إلى ارتفاع معدلات التشخيص الخاطئ للمجموعات الممثلة تمثيلاً ناقصاً، مما يقوض موثوقية وعدالة أداة التشخيص التي تعتمد على الذكاء الاصطناعي.



التأكد من أن بيانات التدريب تشمل نطاقاً واسعاً وتمثيلاً للحالة أو الفئة السكانية المعنية. يتضمن ذلك تنوع نطاق شدة الحالات، والتركيب السكانية للمريض، والخصائص الأخرى ذات الصلة لتعكس التباين الواقعي في الحالة التي تتم دراستها. بالإضافة إلى ذلك، فإن استخدام تقنيات التحقق القوية التي تختبر نموذج الذكاء الاصطناعي عبر مجموعات فرعية مختلفة ومراقبة أدائه بشكل مستمر في الممارسة السريرية يمكن أن يساعد على تحديد وتصحيح تحيز النطاق.



تحيز الوصول التشخيصي

يحدث تحيز الوصول التشخيصي عندما تختلف قدرة الأفراد أو المجموعات على الوصول إلى الخدمات التشخيصية، مما يؤدي إلى تفاوت في فرص التشخيص الدقيق أو في توقيته. ينشأ هذا التحيز بسبب عوامل مثل الموقع الجغرافي، أو الوضع الاجتماعي والاقتصادي، أو التغطية الصحية، ويؤثر في عدالة وجودة الرعاية المدعومة بالذكاء الاصطناعي.

يؤثر تحيز الوصول التشخيصي في دقة وفعالية خدمات الرعاية الصحية المدعومة بالذكاء الاصطناعي، من خلال تفاوت فرص الحصول على التشخيص المناسب بين الأفراد. يؤدي هذا التحيز إلى تأخر في تحديد الحالات أو تشخيصها بشكل غير دقيق، ويسهم في تفاوت جودة الرعاية بين الفئات المختلفة بحسب ظروفهم الجغرافية أو الاجتماعية أو الاقتصادية.



النماذج التي تم تدريبها على صور طبية تأتي من أجهزة تصوير معينة أو من مواقع جغرافية محددة قد لا تعمل بشكل جيد مع صور من أجهزة أخرى أو من مواقع جغرافية مختلفة. هذا يمكن أن يؤدي إلى تشخيصات غير دقيقة للمرضى الذين يتم تصويرهم باستخدام تقنيات مختلفة.



تحليل البيانات لكشف أي تحيزات موجودة قبل تدريب النموذج. يمكن استخدام أدوات تحليل البيانات لاكتشاف الاختلافات في التوزيع وفهم كيفية تأثير هذه التحيزات على النموذج.



تحيز انجراف المفهوم

يحدث عندما تتغير الخصائص الإحصائية للمتغير المستهدف، الذي يتنبأ به النموذج، مع مرور الوقت. يعني هذا الانحراف، بشكل حصري، أن الأنماط والعلاقات التي تعلمها النموذج من بيانات التدريب لم تعد قائمة بنفس الطريقة، مما يؤدي إلى تراجع في أداء النموذج. يمكن أن يكون انحراف المفهوم تدريجياً، أو مفاجئاً، أو دورياً، عاكساً للتغيرات في العملية الأساسية أو البيئة التي يهدف النموذج إلى استيعابها.

يؤثر تحيز انجراف المفهوم في دقة وموثوقية أنظمة الذكاء الاصطناعي في البيئات المتغيرة، حيث يؤدي إلى تراجع أداء النموذج بسبب تغير العلاقات الأساسية التي تم التدريب عليها. دون معالجة هذا الانجراف، تصبح النماذج غير فعالة، مما يؤدي إلى قرارات غير دقيقة وخسائر مالية في التطبيقات الحساسة.



التأثير

في سياق اكتشاف الاحتيال، يمكن أن يحدث انجراف في المفاهيم مع تطور السلوكيات الاحتيالية، فقد يجد نموذج التعلم الآلي الذي تم تدريبه على اكتشاف الاحتيال بناءً على أنماط معاملات معينة لكن مع مرور الوقت، قد يغير المحتالون أساليبهم، ونتيجة لذلك تتضاءل فاعلية النموذج في التعرف على الاحتيال لأن المفهوم (نمط المعاملات الاحتيالية) الذي تعلمه أثناء التدريب لم يعد يتطابق مع الواقع الجديد.



المثال

مراقبة أداء النموذج بشكل مستمر وتحديث النموذج بانتظام ببيانات جديدة تعكس أحدث الأنماط والعلاقات. يمكن أن تكون تقنيات مثل التعلم عبر الإنترنت، حيث يتم تحديث النموذج بشكل تدريجي مع وصول بيانات جديدة، أو أساليب التجميع التي تجمع بين نماذج متعددة للتكيف مع التغيرات. بالإضافة إلى ذلك، يمكن أن يساعد إعداد أنظمة للكشف عن انجراف المفهوم تلقائياً على تحديد الهوية والاستجابة لها في الوقت المناسب.



التخفيف

تحيز انجراف النموذج

يشير انجراف النموذج إلى تدهور تدريجي في أداء نموذج تعلم الآلة مع مرور الوقت، دون وجود سبب واضح وفوري. يحدث هذا الانخفاض نتيجة تغيرات في بيئة التطبيق، مثل تبدل أنماط البيانات المستخدمة في التدريب (انجراف البيانات)، أو تغير العلاقات بين المتغيرات (انجراف المفهوم)، أو انخفاض جودة البيانات المدخلة.

يؤثر تحيز انجراف النموذج في دقة التنبؤات مع مرور الوقت، نتيجة تغيرات في البيانات أو بيئة التطبيق. كما يقلل من فاعلية النماذج في البيئات الحرجة مثل الرعاية الصحية أو مراقبة السلامة، ويؤدي إلى قرارات دون المستوى المطلوب، مما يعرض حياة الأفراد أو كفاءة الأنظمة للخطر.



التأثير

خوارزمية معالجة اللغات الطبيعية المستخدمة لتصفية البريد الإلكتروني العشوائي. تم تدريبها على التعرف على الرسائل غير المرغوب فيها استناداً إلى ميزات محددة مثل "أرباح اليانصيب". مع مرور الوقت، يطور مرسلو البريد العشوائي أساليبهم، ويقدمون أساليب لم يتعرف عليها النموذج. وبالتالي، فإن النموذج الذي تم تدريبه على البيانات منذ سنوات يفشل في التعرف على هذه الأنواع الجديدة من البريد العشوائي، مما يؤدي إلى انخفاض فعاليته.



المثال

تنفيذ أنظمة التعلم التكيفية التي يمكنها تحديث النموذج استجابة للتغيرات المكتشفة، ويمكن أن تساعد إعادة التدريب المنتظم للنموذج باستخدام البيانات الجديدة التي تعكس السياق التشغيلي الحالي على الحفاظ على دقته. بالإضافة إلى ذلك، فإن تطوير آليات للكشف عن الانجراف في وقت مبكر، مثل الاختبارات الإحصائية أو مراقبة الأداء مقابل المعايير، يسمح بالتدخلات في الوقت المناسب.



التخفيف

تحيز انهيار النموذج

يحدث تحيز انهيار النموذج عندما تفقد نماذج التعلم العميق، خصوصاً النماذج التوليدية مثل الشبكات التوليدية التنافسية (GANs) أو الترميز التلقائي المتغير (VAEs)، قدرتها على إنتاج مخرجات متنوعة. يظهر هذا التحيز حين يبدأ النموذج في إنشاء نتائج متشابهة جداً أو متطابقة، مما يحد من الابتكار والتنوع في مخرجاته.

يؤثر تحيز انهيار النموذج في قدرة أنظمة الذكاء الاصطناعي التوليدية على إنتاج مخرجات متنوعة وفعالة. يؤدي إلى تكرار غير مفيد في النتائج، ويقوض الابتكار في النماذج المصممة لإنشاء المحتوى أو تعزيز البيانات، مما يحد من جدواها العملية واتساع استخداماتها عبر السياقات المختلفة.



التأثير

نموذج الشبكات التوليدية التنافسية (GAN) تم تدريبه على إنشاء صور لأزياء متنوعة، إلا أن النموذج يتعلم توليد فساتين سوداء فقط بشكل متكرر. يحدث ذلك بسبب اعتماد النموذج على مجموعة فرعية محدودة من البيانات التي ركزت بشكل غير متوازن على هذا النوع من الملابس، مما يؤدي إلى فقدان قدرته على إنتاج تصاميم متنوعة.



المثال

تنفيذ الاستراتيجيات ضمن بناء النموذج وعملية التدريب لتشجيع التنوع في المخرجات. ويمكن أن تكون تقنيات مثل إدخال الضوضاء إلى المدخلات، أو استخدام طرق التنظيم، أو استخدام بيانات نموذجية مختلفة فعالة وأقل عرضة للانهيار. بالإضافة إلى ذلك، تدريب النموذج على المهام البسيطة قبل الانتقال إلى المهام الأكثر تعقيداً.



التخفيف

تحيز بيئي

يحدث هذا التحيز عندما يتأثر أداء وسلوك نموذج الذكاء الاصطناعي بعوامل في البيئة التي يتم نشره فيها، والتي لم تكن موجودة أو لم يتم أخذها في الاعتبار أثناء مرحلتي التدريب والاختبار. يمكن أن ينتج هذا التحيز عن الاختلافات في تفاعلات المستخدم أو التغيرات الجغرافية أو الظروف البيئية أو البيئة التقنية.

يؤثر التحيز البيئي في دقة أداء أنظمة الذكاء الاصطناعي عند تطبيقها في بيئات تختلف عن تلك التي تم تدريبها فيها. وقد يؤدي إلى قرارات غير موثوقة نتيجة عدم التكيف مع الظروف التقنية أو الجغرافية الجديدة، مما يقلل من كفاءة النظام، ويقوض الثقة المجتمعية، ويضعف جدوى الاعتماد عليه في السياقات المتغيرة.



التأثير

نموذج ذكاء اصطناعي تم تدريبه للتنبؤ بفشل المعدات في أحد المصانع باستخدام بيانات من مجموعة محددة من أجهزة الاستشعار الموجودة على الآلات. وبمرور الوقت، يقوم المصنع بتحديث معداته، واستبدال أجهزة الاستشعار القديمة بأخرى أحدث وأكثر حساسية. وقد تجمع هذه المستشعرات الجديدة البيانات بترددات مختلفة أو بتنسيقات مختلفة. ونتيجة لذلك، تتدهور تنبؤات النموذج لأنه لم يتم تدريبه على مراعاة هذا التغيير في البيئة الفنية لجهاز الاستشعار، مما يؤدي إلى تحيز بيئي.



المثال

استخدام بيانات تدريب متنوعة واختبار وتقييم النموذج في مجموعة واسعة من البيئات، بالإضافة إلى استخدام تقنيات التعلم الآلي المتقدمة ومراجعة وتحسين بيانات التدريب بانتظام، كما يستحسن التفاعل مع المجتمع لضمان التوافق مع احتياجاته المتغيرة.



التخفيف

تحيز تأثير الإجماع الوهمي

يشير تحيز تأثير الإجماع الوهمي إلى ميل أنظمة الذكاء الاصطناعي إلى تبني آراء أو معتقدات تبدو شائعة فقط لأن البيانات التي تم التدريب عليها تعكس قيم ومواقف مبتكريها. يفترض المطورون أحياناً أن وجهات نظرهم تمثل الأغلبية، مما يؤدي إلى استبعاد أو إغفال بيانات تعبر عن تنوع حقيقي، وينتج عن ذلك مخبرات متحيزة تعزز شعوراً زائفاً بالتوافق.

يؤدي تحيز الإجماع الوهمي إلى إنتاج نماذج تعكس رؤى ضيقة يفترض أنها شائعة، مما يهمل وجهات النظر المتنوعة ويقلل من شمولية النظام. يمكن أن يؤدي ذلك إلى استجابات متحيزة تؤثر سلباً في العدالة المجتمعية، وتعزز الممارسات التمييزية، خاصة في المجالات التي تتطلب تمثيلاً عادلاً لكافة الفئات.



التأثير

نظام ذكاء اصطناعي لتصفية محتوى وسائل التواصل تم تدريبه على بيانات تعكس ثقافة واحدة، مما جعله يصنف آراءً ثقافية مختلفة كغير ملائمة. حدث هذا لأن المطورين افترضوا أن ثقافتهم تمثل الإجماع، مما أدى إلى تهمل التنوع الحقيقي في المحتوى.



المثال

جمع الملاحظات بانتظام من مجموعات مختلفة، بما في ذلك عامة الناس، والمتخصصون في الصناعة، وصناع السياسات، والمستخدمون النهائيون، بالإضافة إلى تسهيل المناقشات مع الأفراد من خلفيات مختلفة مما يساعد المطورين والباحثين على سماع وجهات نظر متنوعة.



التخفيف

تحيز تأثير التأطير

يحدث تحيز تأثير التأطير عندما تؤثر طريقة عرض المعلومات في كيفية تفسيرها واتخاذ القرار بناءً عليها. يعتمد هذا التحيز على أن أسلوب صياغة البيانات يمكن أن يغير استجابة النظام، مما يؤدي إلى استنتاجات أو قرارات مختلفة رغم أن المحتوى الأساسي للمعلومة يظل ثابتاً.

يؤثر تحيز التأطير في قرارات النظام من خلال طريقة تقديم البيانات، حيث يؤدي اختلاف الصياغة إلى تباين في التفسير رغم ثبات المحتوى. هذا التحيز قد يوجه استجابات الذكاء الاصطناعي في اتجاهات غير محايدة، مما يعزز التحيزات القائمة ويؤثر سلباً في جودة التوصيات والقرارات النهائية.



في نظام ذكاء اصطناعي لتقديم توصيات علاجية، عند عرض نتائج اختبار دواء معين بصيغة: "يؤدي هذا الدواء إلى تحسن لدى 80% من المرضى"، قد يُفسر ذلك على أنه فعال وآمن. بينما لو عُرضت نفس النتيجة بصيغة: 20% "من المرضى لم يتحسنوا عند استخدام هذا الدواء"، فقد يُقيّم النظام الخيار العلاجي بدرجة أقل إيجابية، رغم أن المعطى الإحصائي ثابت.



ضمان تعرض صانعي القرار لكل من الصياغات الإيجابية والسلبية لنفس المعلومات، وتعزيز التفكير الناقد في طريقة تأطير المعلومات والأخذ في الاعتبار تفسيرات بديلة، كما يشمل استخدام نماذج موحدة أو قوائم مراجعة أو معايير تقلل من التأثير الذاتي الناتج عن طريقة عرض المعلومات.



تحيز تأثير القرون أو تأثير السلبيات

يشير إلى التحيز الذي يحدث عندما يتأثر تقييم الذكاء الاصطناعي لفرد أو كيان بشكل غير متناسب بسمّة سلبية واحدة أو سلوك. يمكن أن يؤدي هذا التحيز إلى تقييم شامل غير عادل بالسلبية، حيث قد يركز الذكاء الاصطناعي، تحت تأثير الجوانب السلبية في بيانات التدريب أو خوارزميات التفسير، بشكل غير مبرر على عيب واحد على حساب صفات أو أداء إيجابي آخر.

يؤدي إلى أحكام غير عادلة وضياع الفرص للأفراد الذين يتمتعون بالكفاءة في مجالات مختلفة. ويقوض هذا التحيز عدالة وفاعلية التقييمات المعتمدة على الذكاء الاصطناعي، مما قد يؤدي إلى الإحباط، وانخفاض التنوع، وعدم الكفاءة داخل المنظمات.



التأثير

أداة التوظيف المعتمدة على الذكاء الاصطناعي يظهر فيها هذا التحيز من خلال احتساب جانب سلبي واحد من طلب التوظيف الخاص بالمرشح بشكل غير متناسب، مثل وجود فجوة في تاريخ التوظيف، على الرغم من أنه يمتلك مؤهلات وخبرات قوية في أماكن أخرى، مما يؤدي ذلك إلى استبعاد المرشحين المحتملين بشكل غير عادل من عملية الاختيار بناءً على عامل منفصل يركز عليه نظام الذكاء الاصطناعي بشكل غير مناسب.



المثال

تدريب أنظمة الذكاء الاصطناعي وتشغيلها وفقاً لمعايير متوازنة تقوم بتقييم مجموعة واسعة من السمات بشكل عادل، دون التركيز غير المبرر على الجوانب السلبية. يمكن أن يشمل ذلك تنفيذ عمليات التحقق داخل الذكاء الاصطناعي لاكتشاف وتصحيح الترجيح السلبي غير المتناسب وتصميم عملية التقييم لتقييم أبعاد الأداء المختلفة بشكل مستقل.



التخفيف

تحيز تأثير الهالة

يحدث عندما تتأثر تقييمات النظام لفرد أو كيان بشكل غير متناسب بسمة إيجابية واحدة بارزة، مما يؤدي إلى تقييم مفرط في الإيجابية لجميع صفاته. يمكن أن يُدخل هذا التحيز من خلال البيانات المستخدمة لتدريب أنظمة الذكاء الاصطناعي، حيث قد يُفترض أن الأفراد الذين لديهم ميزة إيجابية بارزة يتفوقون في جميع المجالات، أو خلال تطبيق الخوارزمية، حيث قد يركز الذكاء الاصطناعي بشكل زائد على جانب إيجابي واحد في عمليات اتخاذ القرار.

يؤثر هذا التحيز في دقة تقييمات أنظمة الذكاء الاصطناعي من خلال تضخيم سمات إيجابية فردية، مما يؤدي إلى تفضيل غير مبرر للأشخاص أو منتجات بناءً على جانب واحد فقط. هذا يخل بالتوازن في صنع القرار، ويتجاهل جوانب الضعف أو مجالات التحسين، مما يحد من عدالة وفاعلية الأنظمة في التقييم الحقيقي والشامل.



التأثير

نظام الذكاء الاصطناعي المصمم للتمتة تقييم أداء الموظفين قد يظهر تأثير الهالة عندما يُخصص تقييم أداء عالٍ بشكل عام بناءً على إنجاز استثنائي لموظف في مشروع واحد فقط. قد يؤدي هذا إلى تجاهل المجالات التي قد يحتاج الموظف إلى تطويرها، مما يؤدي إلى تقييمات متحيزة لا تعكس بدقة ملف أداء الموظف الشامل.



المثال

استخدام معايير تقييم متوازنة ومتعددة الأوجه تقيّم مجموعة واسعة من السمات بشكل مستقل. ويعد تدريب أنظمة الذكاء الاصطناعي على مجموعات البيانات المتنوعة التي تسلط الضوء على مدى تعقيد الأداء، بما في ذلك آليات اكتشاف وتصحيح الوزن غير المتناسب لأي سمة منفردة، بمثابة خطوات أساسية.



التخفيف

تحيز تأثير هوثورن

يحدث هذا التحيز عندما يتلقى النظام بيانات مشوهة خلال مرحلة التدريب بسبب تغير سلوك المستخدمين عند إدراكهم أنهم خاضعون للمراقبة أو التقييم. يؤدي ذلك إلى بناء نماذج ذكاء اصطناعي تستند إلى أنماط غير واقعية من السلوك، مما يضعف من دقة النظام وكفاءته عند التعامل مع المستخدمين في الظروف الطبيعية.

يؤثر تحيز هوثورن في موثوقية البيانات المستخدمة لتدريب نماذج الذكاء الاصطناعي، إذ إنه من الممكن أن يغير الأفراد سلوكهم عند علمهم بأنهم قيد المراقبة. هذا التغير يؤدي إلى تدريب النماذج على بيانات لا تعكس السلوك الطبيعي، مما يضعف دقة التنبؤات أو التوصيات.



التأثير

عند تطوير نظام ذكاء اصطناعي لمراقبة إنتاجية الموظفين، قد يزيد الموظفون من نشاطهم مؤقتاً عندما يدركون أن النظام يتتبع تحركاتهم وساعات عملهم. نتيجة لذلك، يتعلم النظام أنماط عمل غير طبيعية ومؤقتة، مما يؤدي إلى نماذج تقييم أداء لا تعكس السلوك الحقيقي للموظفين في الظروف المعتادة.



المثال

جمع البيانات في بيئات طبيعية حيث لا يكون المستخدمون على علم بأنهم مراقبون، وتنفيذ اختبارات غير معلومة للنظام لتجنب التأثيرات الناتجة عن الوعي بالمراقبة. كما يمكن استخدام دراسات طويلة الأمد ومجموعات تحكم لمراقبة السلوك بمرور الوقت، مع تأكيد إخفاء هوية البيانات المستخدمة وتوظيف أدوات مراقبة آلية لتقليل التدخل البشري المباشر.



التخفيف

تحيز تاريخي

يحدث عندما تشتمل مجموعة البيانات المستخدمة لتدريب نموذج الذكاء الاصطناعي على افتراضات، أو معايير، أو ظروف من الماضي لم تعد دقيقة أو ذات صلة بالسياق الحالي. يمكن لهذا النوع من التحيز أن يتسبب بشكل غير مقصود في استمرار أنظمة الذكاء الاصطناعي في تكريس الصور النمطية القديمة أو اتخاذ قرارات استناداً إلى ظروف قد تطورت.

يؤدي هذا التحيز إلى اعتماد أنظمة الذكاء الاصطناعي على بيانات قديمة لا تعكس السياقات أو الاحتياجات الحالية، مما يتسبب في نتائج غير عادلة أو غير ملائمة. يعزز هذا التحيز الصور النمطية المتجذرة تاريخياً، خاصة في القطاعات التي تتطلب حساسية عالية للتغيرات الاجتماعية كالتوظيف والعدالة والإقراض.



عندما يتم تدريب النموذج على بيانات الإقراض التاريخية التي تعكس الممارسات التمييزية السابقة، مثل الاستبعاد أو معايير الإقراض المتحيزة، قد يستمر حرمان المتقدمين من أحياء أو مجموعات سكانية معينة، على الرغم من حدوث تغييرات في القوانين والأعراف المجتمعية التي تحظر مثل هذا التمييز في الوقت الحالي. وحتى لو كانت ممارسات الإقراض الحالية عادلة، فإن تحيز البيانات التاريخية يمكن أن يجعل الذكاء الاصطناعي يكرر حالات التمييز السابقة.



اتباع نهج استباقي ومتعدد الأوجه، يتضمن ذلك تحديث مجموعات بيانات التدريب بانتظام لتعكس الظروف الحالية، وتحليل سياق جمع البيانات لتحديد التحيزات التاريخية، واستخدام مصادر البيانات المتنوعة لضمان تمثيل أوسع، واستخدام تقنيات تصحيح التحيز مثل إعادة وزن نقاط البيانات، ومراقبة أداء نظام الذكاء الاصطناعي بشكل مستمر.



تحيز تأسيسي

يحدث عندما يتم إدخال معلومات أو بيانات من نتائج نموذج سابق في بيانات التدريب لنموذج لاحق. هذا يمكن أن يؤدي إلى تأثيرات غير دقيقة أو مضللة في النموذج الجديد، حيث يتأثر النموذج بالمعلومات التي ربما تكون قد تم تحريفها أو تضخيمها من النموذج السابق.

يؤدي هذا التحيز إلى تكرار الأخطاء أو التحيزات السابقة عبر إدخال مخرجات النماذج القديمة ضمن بيانات تدريب جديدة، مما يخلق دورة تراكمية من الانحرافات. عندما تُبنى النماذج الحديثة على نتائج مشوهة من نماذج سابقة، تصبح هذه الانحرافات جزءاً من أساس التنبؤ، مما يقلل من دقة الأنظمة ويقوض موثوقيتها على المدى الطويل.



التأثير

نموذج ذكاء اصطناعي للمساعدة على ترشيح الموظفين تم تدريبه على نتائج من نماذج سابقة تحتوي على تحيزات، مما قد يعزز التحيزات القائمة ضد مجموعات معينة، وبالتالي يؤدي ذلك إلى تكرار الأخطاء في قرارات التوظيف وترسيخ أنماط التمييز داخل النظام بدلاً من تصحيحها.



المثال

فصل بيانات التدريب عن بيانات النتائج التي تم الحصول عليها من النماذج السابقة، كما يجب أن تكون البيانات المستخدمة لتقييم النموذج مستقلة تماماً عن البيانات المستخدمة في تدريب النموذج.



التخفيف

تحيز تفاعلي

يشير إلى إدراج تحيزات الإنسان في أنظمة الذكاء الاصطناعي نتيجة تفاعلها مع المستخدمين. يشمل ذلك التحيزات الموجودة في بيانات المستخدم، وطريقة تفاعل المستخدمين مع أنظمة الذكاء الاصطناعي، وآليات التغذية الراجعة التي تُعلم وتُكيف عمليات تعلم الذكاء الاصطناعي.

يؤثر هذا التحيز في دقة وحيادية أنظمة الذكاء الاصطناعي نتيجة تعلمها المستمر من تفاعلات المستخدمين، والتي قد تعكس آراء متحيزة أو أحكاماً مسبقة. يؤدي ذلك إلى تعزيز التحيزات بدلاً من تصحيحها، إذ تبدأ النماذج في تفضيل محتوى معين يعكس تلك السلوكيات البشرية، مما يهدد بالتضليل ويؤثر سلباً في التنوع والعدالة في النتائج المقدمة.



التأثير

في تقنيات التعرف على الصوت، إذا كانت غالبية تفاعلات المستخدم تأتي من ناطقين أصليين للغة أو لهجة معينة، فسيصبح نظام الذكاء الاصطناعي أفضل في فهم أنماط الكلام والاستجابة لها بينما يواجه صعوبة مع اللكنات أو اللهجات الأقل تمثيلاً في تفاعلاته.



المثال

التأكد من تعرض نظام الذكاء الاصطناعي لمجموعة واسعة من تفاعلات المستخدم، مع تنفيذ أدوات وخوارزميات مخصصة لاكتشاف التحيزات وتصحيحها لحظياً، بما يعزز من عدالة أداء النظام وموثوقية نتائجه خلال الاستخدام المستمر.



التخفيف

تحيز توليدي

يشير إلى التحيزات التي تتبناها النماذج التوليدية من بيانات التدريب، مما يؤدي إلى إنتاج مخرجات قد تعكس أنماطاً اجتماعية منحازة أو غير متوازنة. ويمكن أن يظهر هذا التحيز على شكل ما يُعرف بالهيار النمط، حيث ينتج النموذج مجموعة محدودة من الأمثلة.

يؤثر هذا التحيز في جودة وتنوع المخرجات التي تنتجها النماذج التوليدية، حيث قد يؤدي إلى تعزيز الصور النمطية أو تجاهل تمثيل بعض الفئات. في التطبيقات الإبداعية، يمكن أن يعكس محتوى النموذج تحيزات اجتماعية أو ثقافية غير مقصودة، مما يُضعف الشمولية. كما قد ينتج النموذج أمثلة مكررة أو محدودة، ما يقلل من الابتكار وفعالية النماذج في توليد محتوى متنوع.



التأثير

عند تدريب نموذج لغوي على كمية كبيرة من نصوص الإنترنت، مثل المقالات ومنشورات وسائل التواصل الاجتماعي، قد تحتوي بيانات التدريب على محتوى غير متوازن يصور بعض الأجناس في أدوار نمطية محددة. نتيجة لذلك، يتعلم النموذج هذه الأنماط المنحازة. وعند طلب إنشاء نص حول موضوع معين، قد يعيد النموذج إنتاج هذه الصور النمطية التي اكتسبها أثناء التدريب، مما يعزز التحيزات بدلاً من تجنبها.



المثال

اتباع نهج يتضمن استخدام بيانات التدريب المتنوعة لتغطية مجموعة واسعة من المواضيع، وتنفيذ تقنيات الكشف عن التحيز وتصحيحه للتخفيف من التحيز في بيانات التدريب ومخرجاته، ووضع مبادئ توجيهية أخلاقية والإشراف عليها لضمان العدالة واعتبارات التأثير المجتمعي، وإشراكها في التقييم المستمر والتعليقات من مختلف أصحاب المصلحة لتحديد التحيز وتصحيحه.



التخفيف

تحيز حجم العينة

يحدث عندما يكون حجم مجموعة البيانات المستخدمة لتدريب أو اختبار نموذج الذكاء الاصطناعي غير كافٍ أو كبير بشكل غير متناسب، مما يؤثر في قدرة النموذج على التعلم، أو التعميم، أو التحقق من النتائج بدقة. يمكن أن يؤدي حجم العينة الصغير بشكل غير مناسب إلى الإفراط في التخصيص وضعف التعميم على بيانات جديدة، بينما قد يتضمن حجم العينة الكبير جداً ضوضاء أو تشويشاً يقلل من أداء النموذج أو تعقيد عملية التعلم بشكل غير ضروري.

يؤدي هذا التحيز إلى إضعاف دقة النموذج وموثوقيته بسبب استخدام مجموعة بيانات صغيرة جداً أو كبيرة بشكل غير مناسب. قد يمنع حجم العينة الصغير النموذج من تعلم الأنماط المهمة، بينما حجم العينة الكبير جداً قد يضيف ضوضاء غير ضرورية. كلا الحجمين يعيق قدرة النموذج على التعميم، مما يؤدي إلى قرارات غير دقيقة أو إهدار للموارد في البيئات الحقيقية.



التأثير

في تحليل الصور الطبية، قد يؤدي استخدام مجموعة بيانات تحتوي على عدد قليل جداً من الصور إلى تفويت نموذج الذكاء الاصطناعي للتغيرات الحرجة في الأمراض، مما يؤدي إلى تشخيصات غير صحيحة. من ناحية أخرى، قد تتضمن مجموعة البيانات الكبيرة جداً عدداً من الخصائص المختلفة، مما يجعل من الصعب على النموذج تحديد المؤشرات الرئيسية لحالات طبية محددة مما يؤدي إلى تقليل دقة التشخيص.



المثال

تقدير الحد الأدنى لحجم العينة اللازم لتحقيق نتائج موثوقة، واستخدام تقنيات التحقق المتبادل لاختبار أداء النموذج على مجموعات فرعية مختلفة من البيانات، ومراقبة أداء النموذج بشكل مستمر لضبط حجم مجموعة البيانات حسب الحاجة. إن تحقيق التوازن بين مجموعة البيانات للتأكد من أنها شاملة بما يكفي لالتقاط الأنماط ذات الصلة وخالية من الضوضاء غير الضرورية أمر بالغ الأهمية لتحسين نماذج الذكاء الاصطناعي.



التخفيف

تحيز دلالي

يشير التحيز الدلالي إلى تشويه المعنى أو التفسير في أنظمة الذكاء الاصطناعي بسبب سياق أو دلالات البيانات. يمكن أن يؤدي إلى تحريفات أو عدم دقة في مهام معالجة اللغة مثل تحليل المشاعر أو تصنيف النص. ويرتبط هذا النوع من التحيز بشكل خاص بمعالجة اللغات الطبيعية (Natural Language Processing).

يؤثر هذا التحيز في دقة أنظمة الذكاء الاصطناعي في فهم اللغة ومعالجة النصوص، مما قد يؤدي إلى نتائج متحيزة أو غير دقيقة عند تحليل المشاعر أو تصنيف المحتوى. كما أنه يعزز الصور النمطية، ويؤثر في عدالة الخوارزميات المستخدمة في التوظيف أو الإشراف الرقمي، ويضعف الشفافية في التطبيقات اللغوية المختلفة.



التأثير

عندما يتم تدريب نموذج لغوي على بيانات نصية تعكس بشكل كبير الصور النمطية لجنس بعض الوظائف. فعلى سبيل المثال عندما يرى النموذج بشكل متكرر عبارات مثل "هو طبيب" أو "هي ممرضة"، فقد يتعلم ربط "الطبيب" بشكل أكبر بالرجال و"الممرضة" بشكل أكبر بالنساء. ونتيجة لذلك، عندما يُطلب منه إنشاء نص أو إكمال جملة، فقد يستمر الذكاء الاصطناعي في إعادة إنتاج هذه الأنماط المتحيزة، حتى في الحالات التي لا يكون فيها الجنس مهماً.



المثال

تطوير وتنفيذ خوارزميات لاكتشاف التحيز وتقليله في تضمين الكلمات ونماذج معالجة اللغات الطبيعية الأخرى. قد يستلزم ذلك تغيير تمثيلات النموذج لتحديد الارتباطات بين مصطلحات محددة ومفاهيم متحيزة. بالإضافة إلى التأكد من أن البيانات المستخدمة لتدريب نماذج الذكاء الاصطناعي متنوعة ومتوازنة وتمثل مجموعة متنوعة من وجهات النظر. وهذا يعني تضمين المعلومات التي تتحدى الافتراضات والأحكام المسبقة الحالية بشكل فعال.



التخفيف

تحيز رقمي

يشير التحيز الرقمي إلى الانحرافات المنهجية التي تنشأ داخل أنظمة الذكاء الاصطناعي نتيجة تحيزات كامنة في البيانات أو الخوارزميات أو قرارات التصميم. تؤثر هذه التحيزات في كيفية معالجة المعلومات وعرضها، وقد تؤدي إلى نتائج غير دقيقة أو غير عادلة تعكس تحيزات ضمنية في بنية النظام نفسه.

يؤدي هذا التحيز إلى قرارات غير عادلة داخل أنظمة الذكاء الاصطناعي، نتيجة الخوارزميات أو البيانات التي تعكس تحيزات بيانية. يمكن أن يُقصي مجموعات سكانية معينة من الحصول على فرص متساوية، مثل رفض مرشحين مؤهلين لأسباب غير مرتبطة بالكفاءة، مما يعزز التمييز الاجتماعي ويقوض ثقة المستخدمين في أنظمة الذكاء الاصطناعي.



التأثير

نموذج ذكاء اصطناعي لتقييم الجدارة الائتمانية تم تدريبه على بيانات مصرفية تاريخية تغيب عنها معلومات كافية عن فئات من أصحاب الدخل غير الثابت، مثل العاملين لحسابهم الخاص. نتيجة لذلك، يُقلل النظام من درجاتهم الائتمانية تلقائياً، رغم التزامهم بالسداد، مما يؤدي إلى حرمانهم من فرص التمويل وزيادة الفجوة الاقتصادية.



المثال

استخدام مجموعات بيانات متنوعة وشاملة لتدريب نماذج الذكاء الاصطناعي، لضمان تضمين جميع الفئات الاجتماعية والخلفيات الثقافية. كما يجب إجراء مراجعات دورية للأنظمة الخوارزمية لكشف وتصحيح أي تحيزات غير مقصودة. بالإضافة إلى ذلك، الشفافية والمساءلة، حيث يجب على الشركات التي تطوّر وتستخدم هذه الأنظمة أن تظل شفافة حول كيفية تدريب نماذجها وضمان وجود آليات للمساءلة في حال اكتشاف أي تحيزات.



التخفيف

تحيز زمني

يشير إلى التحيزات التي تظهر بسبب التغيرات في السكان والسلوكيات عبر الزمن. يحدث هذا النوع من التحيز عندما تفشل النماذج المدربة على بيانات من فترة زمنية معينة في مراعاة أو التنبؤ بدقة بالسلوكيات والاتجاهات التي تتطور أو تظهر في فترات لاحقة. الطبيعة الديناميكية للمجتمعات البشرية وتطور اتجاهاتها يمكن أن تؤثر بشكل كبير في صلة ودقة تنبؤات وتحليلات الذكاء الاصطناعي.

يؤثر هذا التحيز في موثوقية النماذج التي تعتمد على بيانات تاريخية لتوقع المستقبل، حيث تصبح التنبؤات غير دقيقة أو غير ملائمة للسياق الحالي. قد تنتج قرارات غير محدثة تعزز عدم المساواة السابقة، وتضعف من دقة التحليلات المستخدمة في مجالات مثل السياسات العامة أو توقعات السوق أو تخطيط الخدمات.



في منصات التواصل الاجتماعي، قد يستخدم المستخدمون وسماً معيناً في بداية النقاش حول موضوع ما، لكن مع مرور الوقت، قد يتغير نمط التفاعل ويستمر النقاش دون استخدام ذلك الوسم، إما لتغير السياق أو لانخفاض الحاجة إليه. إذا تم تدريب نموذج ذكاء اصطناعي على البيانات المبكرة فقط التي تعتمد على الوسم، فقد يُغفل الجزء الأكبر من النقاشات اللاحقة، مما يؤدي إلى تحليل غير مكتمل أو منحاز حول مدى انتشار الموضوع أو أهميته.



دمج الآليات التي تسمح لأنظمة الذكاء الاصطناعي بالتكيف مع أنماط البيانات المتغيرة بمرور الوقت. يمكن أن يشمل ذلك تقنيات مثل التعلم المستمر، حيث يتم تحديث النماذج بانتظام ببيانات جديدة، واستخدام النماذج الحساسة للوقت والتي تأخذ في الاعتبار الديناميكيات الزمنية للبيانات. بالإضافة إلى ذلك، فإن دمج مصادر البيانات المتنوعة والنظر في سياق وتوقيت جمع البيانات يمكن أن يساعد على تحديد وتقليل تأثير التغيرات الزمنية.



تحيز سياقي

يحدث عندما يؤثر سياق جمع البيانات، أو معالجتها، أو تفسيرها في نتائج أنظمة الذكاء الاصطناعي، مما يؤدي إلى نتائج منحرفة أو متحيزة. يعكس هذا التحيز الشروط أو الظروف المحيطة بإنشاء واستخدام البيانات، التي قد تشكل بشكل غير مقصود فهم وقرارات الذكاء الاصطناعي. ويمكن أن ينبع التحيز السياقي من الوقت، أو المكان، أو الشروط الخاصة التي تم فيها جمع البيانات، مؤثراً في مدى دقة تطبيق نظام الذكاء الاصطناعي.

يؤدي هذا التحيز إلى تقليل قدرة أنظمة الذكاء الاصطناعي على التعميم خارج السياق الذي تم تدريبها عليه. فقد تنتج النماذج توصيات أو تنبؤات غير دقيقة عند استخدامها في ظروف جديدة. يؤثر ذلك سلباً في التطبيقات الواقعية، ويحد من فاعلية الذكاء الاصطناعي في البيئات المتغيرة أو المتعددة الثقافات.



التأثير

في أنظمة التعرف على الصور، يتم التدريب في الغالب على صور من مواقع جغرافية محددة، مثل البيئات الحضرية في البلدان المتقدمة. وعندما يتم تطبيق هذه الأنظمة على صور من المناطق الريفية أو البلدان النامية، فقد ينخفض معدل الدقة بشكل كبير لأن النماذج لم تتعلم من مجموعة متنوعة بما فيه الكفاية من السياقات، مما يؤدي إلى سوء فهم للإشارات أو العناصر البيئية المختلفة.



المثال

ضمان تنوع البيانات عبر سياقات مختلفة، بما في ذلك الأوقات والمواقع والظروف المختلفة التي يتم بموجبها جمع البيانات. قد يتطلب ذلك البحث عن البيانات التي تغطي نطاقاً أوسع من السيناريوهات مما تم تضمينه في البداية أو توليدها. بالإضافة إلى ذلك، يعد تطوير أنظمة الذكاء الاصطناعي مع التركيز على قابلية التعميم والقوة في سياقات مختلفة أمراً بالغ الأهمية.



التخفيف

تحيز ضد الأمراض غير المقبولة

يمثل التحيز ضد الأمراض غير المقبولة نمطاً من الانحراف في أنظمة الذكاء الاصطناعي يظهر عندما تُعامل بعض الحالات الصحية أو الأمراض بحساسية أقل أو بتمييز ضمني، نتيجة تصورات اجتماعية أو أخلاقية سلبية عنها. يؤدي هذا التحيز إلى تقليل جودة الرعاية أو تجاهل احتياجات المرضى الذين يعانون من هذه الحالات.

يؤثر هذا التحيز في عدالة أنظمة الرعاية الصحية المدعومة بالذكاء الاصطناعي، حيث قد تقلل من شأن احتياجات مرضى يعانون من أمراض ذات وصمة اجتماعية. مما يؤدي إلى إغفال الأعراض، أو تأخير العلاج، أو تجاهل البحث الطبي في تلك المجالات، وقد يُفاقم ذلك معاناة المرضى ويؤثر في تكافؤ فرص الرعاية.



التأثير

في نظام ذكاء اصطناعي يُستخدم لتصنيف أولويات المرضى في الطوارئ، قد يُدرج مرضى اضطرابات الأكل أو الإدمان في فئات أقل أولوية بناءً على بيانات تدريبية تعكس تحيزات اجتماعية تجاه هذه الحالات. رغم أن أعراضهم قد تكون حرجة طبياً، فإن النظام قد يُقلّل من شأنها تلقائياً. هذا التحيز يؤدي إلى قرارات غير منصفة، مثل تأخير العلاج أو رفض الإدخال للمستشفى، مما يؤثر سلباً في العدالة وجودة الرعاية الصحية.



المثال

استخدام بيانات تدريب شاملة ومتنوعة، مع إجراء تدقيق دوري للخوارزميات بهدف تحديد التحيزات ومعالجتها، بالإضافة إلى تعزيز الوعي والتعليم بين المهندسين والمستخدمين لدعم تطوير أنظمة أكثر إنصافاً ودقة في الأداء.



التخفيف

تحيز ضمني

يشير هذا التحيز إلى التحيزات اللاواعية التي تنتقل إلى أنظمة الذكاء الاصطناعي عبر البيانات والخوارزميات. رغم أن الأنظمة لا ترمج صراحة على التمييز، إلا أنها تتعلم من الأنماط المنحازة في البيانات التاريخية، مما يؤدي إلى مخرجات تفضل مجموعات معينة دون قصد، نتيجةً لتحيزات مصممي الأنظمة أو القائمين على جمع البيانات.

يؤثر هذا التحيز في عدالة القرارات الصادرة عن أنظمة الذكاء الاصطناعي، من خلال ترسيخ أحكام مسبقة ناتجة عن تحيزات لا واعية في البيانات والتصميم. يمكن أن يؤدي إلى تمييز ضد فئات محددة، ويُعزز الفجوات الاجتماعية، ويقلل من ثقة المستخدمين في الأنظمة الذكية، خاصة في مجالات حساسة مثل التوظيف والرعاية الصحية والإقراض.



التأثير

في أدوات التوظيف المعتمدة على الذكاء الاصطناعي. إذا كانت بيانات التدريب لهذه الأدوات تتكون من سير ذاتية ناجحة من مجموعة سكانية لا تمثل كامل السكان الباحثين عن عمل، فقد يتعلم الذكاء الاصطناعي ضمناً تفضيل المتقدمين الذين يتناسبون مع هذه السير الذاتية، مما سيضر الفئات الأخرى التي لم يتم تمثيلها.



المثال

رفع مستوى الوعي وتوفير التعليم حول التحيزات الضمنية لأولئك المشاركين في تطوير الذكاء الاصطناعي، وضمان جمع بيانات متنوعة وتمثيلية، وتنفيذ تقنيات الكشف عن التحيز وتصحيحه، ومراجعة واختبار أنظمة الذكاء الاصطناعي بشكل منتظم بعد النشر بحثاً عن التحيزات الناشئة.



التخفيف

تحيز عدم الاستجابة

يشير إلى التشوهات في مجموعة البيانات المستخدمة لتدريب نماذج الذكاء الاصطناعي أو التحقق منها أو اختبارها التي تنشأ عندما تفشل نسبة كبيرة من البيانات المستهدفة في الرد أو المشاركة في عملية جمع البيانات. يحدث هذا التحيز إذا كان عدم الاستجابة منهجياً، بمعنى أن أنواعاً معينة من البيانات أو المجموعات أكثر عرضة للغياب عن مجموعة البيانات من غيرها. إذا لم يكن عدم الاستجابة عشوائياً ويتبع نمطاً، يمكن أن يؤدي ذلك إلى مجموعة بيانات متحيزة.

يؤدي هذا التحيز إلى انخفاض تمثيل بعض الفئات في البيانات التدريبية، مما يقلل من قدرة النماذج على التعميم ويزيد احتمالية إصدار قرارات متحيزة. تظهر آثاره بوضوح عندما تفشل الأنظمة في التعامل مع مجموعات لم تشارك بفاعلية في جمع البيانات، وهو ما يقاوم فجوة العدالة ويُضعف فاعلية النظام.



نظام للذكاء الاصطناعي يهدف إلى فهم رضا المرضى عن خدمات الرعاية الصحية بناءً على بيانات الاستطلاع. إذا تلقى الاستطلاع ردوداً من المرضى الذين لديهم تجارب إيجابية، في حين أن الذين لديهم تجارب سلبية يكونون أقل نسبة مشاركة، فإن مجموعة البيانات الناتجة سوف تنحرف نحو التعليقات الإيجابية، وقد يستنتج نموذج الذكاء الاصطناعي الذي تم تدريبه على هذه البيانات أن رضا المرضى مرتفع بشكل عام، متجاهلاً التعليقات النقدية من المرضى غير الراضين الذين لم يستجيبوا للاستطلاع.



تحسين جودة بيانات التدريب في أنظمة الذكاء الاصطناعي وذلك بمتابعة غير المستجيبين لفهم ومعالجة أسباب عدم المشاركة، وضمان أن أدوات جمع البيانات مثل الاستطلاعات تكون مختصرة ومباشرة لتعزيز معدلات الإكمال. كما ينبغي تحديد الجمهور المستهدف بوضوح ووضع أهداف دقيقة لعملية جمع البيانات لتوجيه اختيار العينات وبناء مجموعات تدريب أكثر تمثيلاً ودقة.



تحيز عدم القبول

يحدث تحيز عدم القبول عندما تُستبعد موضوعات أو قضايا مثيرة للجدل من مراحل البحث والتطوير في الذكاء الاصطناعي، بسبب توجهات الباحثين أو الجهات الممولة. يؤدي ذلك إلى تدريب نماذج ذكاء اصطناعي غير شاملة، تفتقر إلى معالجة قضايا اجتماعية أو ثقافية مهمة، مما يقيد عدالة الأنظمة ويحد من فعاليتها في البيئات المتنوعة.

يؤثر هذا التحيز في شمولية البحث والتطوير في مجال الذكاء الاصطناعي، حيث يؤدي إلى استبعاد قضايا حساسة أو مثيرة للجدل من نطاق الدراسة. قد ينتج عنه فهم ناقص للواقع الاجتماعي والثقافي، ويحد من تطور حلول عادلة ومستجيبة لتنوع البيئات والسياقات.



التأثير

إذا كان هناك مشروع لتدريب نموذج ذكاء اصطناعي على تحليل العوامل المؤثرة في الأداء الأكاديمي للطلاب، قد يتجنب الباحثون تضمين بيانات حول تأثير الصحة النفسية بسبب حساسيتها الاجتماعية. نتيجة لذلك، يتعلم النموذج أن يتجاهل هذا العامل المهم، مما يؤدي إلى تحليلات ناقصة لا تعكس جميع جوانب الأداء الأكاديمي الحقيقي.



المثال

تعزيز الوعي الثقافي والاجتماعي لدى الباحثين والممولين حول أهمية دراسة المواضيع المحتملة للتحيز. بالإضافة إلى تشجيع البحث على الاستقلالية والشجاعة في التعامل مع المواضيع المعقدة والمثيرة للجدل وتعزيز استخدام تصاميم الدراسات التي تسمح بتجنب الانحياز والحفاظ على الشفافية. كما يمكن تعزيز مراجعة الأقران والتحليل المستقل للبيانات لضمان موضوعية النتائج. كذلك تعزيز التواصل المفتوح والشفاف حول القرارات المتعلقة بالمواضيع والأساليب المختارة للبحث.



التخفيف

تحيز عدم المتانة

يشير تحيز عدم المتانة إلى ضعف قدرة نموذج الذكاء الاصطناعي على التعامل مع تغيرات بسيطة في البيانات المدخلة، مثل الأخطاء الإملائية أو تغير تنسيق النص. تؤدي هذه الحساسية المفرطة إلى تغيرات كبيرة في النتائج أو التنبؤات، مما يقلل من موثوقية النموذج ويُضعف أدائه في البيئات الواقعية والمتنوعة.

يُضعف هذا التحيز أداء النماذج الذكية عند التعامل مع تغييرات طفيفة في البيانات، مما يقلل من دقة التنبؤات واستقرار النتائج. ويكون تأثير هذا التحيز ملحوظاً في التطبيقات الحساسة مثل القيادة الذاتية أو التشخيص الطبي، حيث قد تؤدي تفاصيل صغيرة إلى قرارات خاطئة ونتائج عالية المخاطر.



التأثير

عندما يكون نظام الذكاء الاصطناعي مدرباً على التعرف على إشارات التوقف والأداء بدقة عالية تحت الظروف الطبيعية. ومع ذلك، إذا تسببت تغييرات طفيفة في مظهر إشارة التوقف (مثل رسومات جرافيتية صغيرة أو ملصقات) في أن يقوم النظام بتصنيفها بشكل خاطئ كإشارة مرور مختلفة، فهذا يشير إلى وجود عدم استقرار. مثل هذا الفشل قد يكون له تداعيات خطيرة على اتخاذ القرارات والسلامة على الطريق للمركبة.



المثال

زيادة تنوع مجموعة بيانات التدريب عبر تقنيات التعزيز -على سبيل المثال، إدخال الضوضاء وإجراء التحويلات- لمساعدة النموذج على تعلم كيفية تحديد المدخلات وتصنيفها بدقة في ظل مجموعة متنوعة من السيناريوهات. علاوة على ذلك، فإن تضمين الحالات المضللة في عملية التدريب قد يؤدي إلى تحسين مرونة النموذج من خلال تعليمه بشكل واضح كيفية اكتشاف المدخلات المشوشة وتصنيفها. وأخيراً، فإن اختبار النموذج بدقة ضد مجموعة متنوعة من الاضطرابات والحالات المضللة يمكن أن يساعد على تحديد نقاط الضعف.



التخفيف

تحيز عدم المخاطرة

يحدث تحيز عدم المخاطرة عندما يُصمَّم أو يُدار نظام الذكاء الاصطناعي بطريقة تميل إلى تفضيل الحلول التي تستهدف القضاء التام على خطر محدد، حتى وإن كانت أقل كفاءة في تقليل المخاطر الكلية. يؤدي هذا التحيز إلى بناء أنظمة تركز على جوانب ضيقة من الأمان أو الأداء، مع إهمال حلول أكثر شمولية تحقق توازناً أفضل بين مختلف أنواع المخاطر.

يؤثر هذا التحيز في كيفية تصميم أنظمة الذكاء الاصطناعي واستراتيجيات السلامة المرتبطة بها، من خلال التركيز المفرط على تجنب نوع واحد من المخاطر بشكل كامل. ويمكن أن يؤدي ذلك إلى تجاهل حلول أكثر شمولاً وإهدار للموارد، مما يحد من قدرة النظام على التعامل مع مجموعة أوسع من المخاطر بفاعلية متوازنة.



التأثير

قد يتأخر مشروع السيارات ذاتية القيادة لسنوات حتى يصل إلى نقطة يتوقع فيها أن يكون معدل الحوادث للذكاء الاصطناعي "صفرًا"، على الرغم من أن تقليل الحوادث بنسبة 90% من خلال الذكاء الاصطناعي يمكن أن يحسن بالفعل السلامة على الطرق بشكل كبير مقارنة بالسائقين البشر.



المثال

تحويل التركيز من القضاء التام على المخاطر إلى الحد منها وإدارتها. وهذا يتطلب تطوير آليات أفضل للشفافية في نماذج الذكاء الاصطناعي، وتحسين أطر الحوكمة. إن تثقيف الناس حول فوائد وقيود الذكاء الاصطناعي يمكن أن يساعد على الحد من المقاومة المبنية على الخوف والسماح باتخاذ قرارات أكثر توازناً. وبدلاً من الدفع نحو معايير خالية من المخاطر، ينبغي للهيئات التنظيمية أن تركز على الأطر التي توازن بين الابتكار والضمانات الأخلاقية، مما يسمح لأنظمة الذكاء الاصطناعي بالنمو بشكل مسؤول.



التخفيف

تحيز عنصري

يجدث عندما تُظهر أنظمة الذكاء الاصطناعي تحيزاً لصالح أو ضد أفراد أو مجموعات بناءً على عرقهم. ينشأ هذا النوع من التحيزات عادةً من بيانات تدريب غير ممثلة، أو خوارزميات معيبة، أو تحيزات كامنة لدى الأشخاص الذين ينشؤون وينشرون هذه الأنظمة. يمكن أن يتجلى ذلك في تطبيقات مختلفة.

يؤثر هذا التحيز بشكل مباشر في عدالة أنظمة الذكاء الاصطناعي، حيث يؤدي إلى قرارات تمييزية ضد مجموعات عرقية معينة نتيجة بيانات غير ممثلة أو تصميمات منحازة. يفاقم الفجوات الاجتماعية، ويقوض الثقة بالنظام، ويضر بالمجتمعات المتضررة من خلال إعادة إنتاج صور نمطية وتفضيلات غير عادلة.



التأثير

نظام الذكاء الاصطناعي للتعرف على الوجوه، حيث تسجل بعض هذه الأنظمة معدلات خطأ أعلى عند التعرف على الأفراد من خلفيات عرقية معينة مقارنة بغيرهم. هذا التفاوت قد يؤدي إلى أخطاء في تحديد الهوية، واعتقالات غير مشروعة، وضعف المساءلة في ممارسات المراقبة، مما يؤثر بشكل غير متناسب في مجتمعات الأقليات.



المثال

جمع واستخدام مجموعات بيانات تمثيلية متنوعة تعكس بدقة التنوع العرقي للسكان الذين يخدمهم نظام الذكاء الاصطناعي. كما أنه من الضروري إشراك أصحاب المصلحة من خلفيات عرقية متنوعة في مرحلتي التطوير والاختبار لتحديد التحيزات المحتملة وفهم آثارها. بالإضافة إلى ذلك، فإن تنفيذ تقنيات العدالة الخوارزمية، مثل خوارزميات الكشف عن التحيز وقيود العدالة، يساعد على تصحيح التحيز العنصري وتقليله في تنبؤات النماذج.



التخفيف

تحيز غياب التعمية

يشير إلى عدم وجود آلية تمنع الباحثين أو المطورين من التعرف على البيانات أو الخوارزميات أو النتائج التي يتم تقييمها أو استخدامها. يمكن أن يؤدي ذلك إلى عدة أنواع من التحيزات ويؤثر في نزاهة وفاعلية أنظمة الذكاء الاصطناعي.

يؤدي هذا التحيز إلى نتائج غير موضوعية نتيجة تأثير المعرفة المسبقة بالمرجات أو البيانات من قبل الباحثين أو المطورين. ويمكن أن يؤثر هذا سلباً في مصداقية التقييم ويزيد من فرص تكرار الأخطاء وتضليل النتائج، مما يقلل من فاعلية النموذج وقدرته على التعميم بدقة.



التأثير

في نموذج تصنيف للصور إذا كان المطورون يعرفون مسبقاً محتوى الصور التي يستخدمونها لاختبار النموذج، فقد يقومون بشكل غير متعمد بتعديل النموذج لتحسين أدائه على هذه الصور المحددة فقط. هذا قد يؤدي إلى تحسين النتائج على مجموعة بيانات معينة ولكن لا يعكس أداء النموذج على بيانات جديدة أو غير معروفة. نتيجة لذلك، قد يكون النموذج أقل قدرة على التعميم ويواجه صعوبات عند التعامل مع بيانات حقيقية أو متنوعة.



المثال

استخدام استراتيجيات مثل فصل الأدوار بين تطوير وتقييم النموذج، وتطبيق التعمية أثناء تقييم البيانات، واستخدام مجموعات بيانات اختبار مستقلة، والاستعانة بمراجعات خارجية. هذه الإجراءات تساعد على ضمان تقييم عادل وموضوعي للنماذج وتفادي التحيزات.



التخفيف

تحيز فرط التخصيص

يحدث عندما يتعلم نموذج الذكاء الاصطناعي بيانات التدريب جيداً، ملتقظاً الضوضاء أو التقلبات العشوائية في البيانات كما لو كانت أنماطاً ذات معنى. مما يُنتج نموذجاً يكون أدائه ممتازاً على مجموعة بيانات التدريب الخاصة به ولكن سيئاً على أي بيانات جديدة غير مرئية. الإفراط في التخصيص هو في الأساس شكل من أشكال التحيز حيث يصبح النموذج متحيزاً بشكل مفرط نحو تفاصيل محددة من بيانات التدريب، على حساب قدرته على التعميم على بيانات العالم الحقيقي الأوسع.

يؤثر هذا التحيز في قدرة النموذج على التعميم، إذ يؤدي إلى أداء ممتاز على بيانات التدريب، وضعيف على البيانات الجديدة. ينتج عن ذلك مخرجات قد تبدو دقيقة لكنها مضللة، مما يؤدي إلى قرارات غير فعالة أو حتى ضارة عند تطبيقها في سياقات العالم الحقيقي.



التأثير

نظام الذكاء الاصطناعي للكشف عن البريد العشوائي، بافتراض أن النموذج قد تم تدريبه على مجموعة من بيانات البريد الإلكتروني، بما في ذلك العبارات المعينة التي تصادف وجودها في رسائل البريد الإلكتروني العشوائية الخاصة بالمجموعة التي تم التدريب عليها ولكنها لا تشير إلى البريد العشوائي. قد يتعلم النموذج كيفية تحديد البريد العشوائي بناءً على الخصائص غير ذات الصلة، ويحقق أداءً جيداً في اختبارات التدريب ولكنه يفشل في تحديد البريد العشوائي بشكل صحيح في الرسائل الواقعية التي لا تشترك في تلك الخصائص المحددة.



المثال

تطبيق بعض الاستراتيجيات مثل تبسيط النموذج لتقليل التعقيد وتجنب التقاط الضوضاء، واستخدام التحقق المتبادل لإجراء تقييم واقعي لقابلية تعميم النموذج، وتطبيق تقنيات التنظيم لتلافي التعقيد، وزيادة حجم مجموعة البيانات وتنوعها وتنفيذ الإيقاف المبكر بناءً على أداء مجموعة التحقق لمنع النموذج من تعلم الضوضاء.



التخفيف

تحيز فرط التعميم

يحدث عندما يكون نموذج تعلم الآلة بسيطاً جداً بحيث لا يستطيع التقاط البنية الكامنة للبيانات التي تم تدريبه عليها. يمكن أن تنشأ هذه البساطة من وجود عدد قليل جداً من المتغيرات في النموذج، أو التدريب غير الكافي، أو الافتراضات المبسطة جداً عن توزيع البيانات. نتيجة لذلك، يفشل النموذج في التعلم بشكل كافٍ من بيانات التدريب ويكون أداؤه ضعيفاً على مجموعة التدريب وعلى البيانات الجديدة غير المرئية، مما يشير إلى أنه لم يلتقط الأنماط الأساسية اللازمة للتنبؤات أو التصنيفات الدقيقة.

يؤدي هذا التحيز إلى أداء ضعيف للنموذج بسبب بساطته المفرطة وفشله في التقاط الأنماط الدقيقة في البيانات. وقد يكون النموذج غير قادر على معالجة التعقيد الفعلي للبيئة المستهدفة، مما يقلل من دقته وكفاءته في التطبيقات الواقعية ويحد من موثوقية نتائجه.



التأثير

نظام الذكاء الاصطناعي المصمم للتنبؤ بأسعار العقارات بناءً على ميزات مثل الموقع والحجم، إذا كان النموذج المستخدم للتنبؤات مفرطاً في التبسيط، وربما يفترض وجود علاقة خطية بين الحجم والسعر فقط دون النظر إلى الموقع أو الحالة؛ فقد يفشل في التنبؤ بدقة بالأسعار في الأسواق حيث يؤثر الموقع والحالة بشكل كبير في قيمة العقار، ويؤدي هذا النقص إلى نموذج غير مفيد للمشتريين أو البائعين أو المتخصصين في العقارات الذين يبحثون عن رؤية دقيقة للسوق.



المثال

زيادة تعقيد النموذج من خلال مزيد من المتغيرات أو الطبقات، وتعزيز هندسة الميزات لاستخراج معلومات أكثر تفصيلاً من البيانات، وضبط التنظيم لتحقيق التوازن بين منع الإفراط في التخصيص والسماح بالتعلم الكافي للنماذج، وتمديد مدة التدريب أو زيادة عدد التكرارات في التدريب لتزويد النموذج بمزيد من الفرص للتعلم من بيانات التدريب بشكل فعال.



التخفيف

تحيز فصلي

يشير هذا التحيز إلى فشل أنظمة الذكاء الاصطناعي في التعامل مع التغيرات الموسمية المنتظمة عند تفسير البيانات أو تقديم التنبؤات. يحدث عندما تُدرَّب النماذج على بيانات تُمثل مواسم معينة بشكل غير متوازن، مما يجعل أداء النظام جيداً فقط في الفترات المشمولة بالتدريب وضعيفاً خارجها، ويؤدي إلى تحليلات وتوقعات غير دقيقة في السياقات الزمنية المختلفة.

يؤثر هذا التحيز في دقة تنبؤات النماذج الذكية عندما تعتمد على بيانات موسمية غير ممثلة لبقية السنة. وقد يؤدي إلى قرارات غير مناسبة في غير أوقات الذروة، ويُسبب سوء تخطيط أو توزيع للموارد، خاصة في القطاعات التي تتأثر بشكل كبير بدورات الموسم.



في مجال الرعاية الصحية، قد يُظهر نموذج الذكاء الاصطناعي الذي تم تطويره للتنبؤ بحالات دخول المستشفى بناءً على البيانات التاريخية تحيزاً موسمياً إذ تتضمن بشكل غير متناسب بيانات من موسم الإنفلونزا، ويؤدي ذلك إلى المبالغة في تقدير حالات دخول المستشفى في أوقات أخرى من العام، مما قد يؤدي إلى مشاكل في التوظيف وتخصيص الموارد ويؤثر في جاهزية المستشفى ورعاية المرضى خلال الأشهر الأقل ازدحاماً.



ضمان تدريب نماذج الذكاء الاصطناعي على البيانات التي تمثل جميع الفصول أو الفترات ذوات الصلة بالتطبيق. ويمكن تحقيق ذلك من خلال دمج العوامل الموسمية في تصميم النموذج، مما يسمح للذكاء الاصطناعي بالتعرف على التغيرات الموسمية في البيانات وتعديلها. بالإضافة إلى ذلك، فإن استخدام تقنيات مثل التعديل الموسمي في المعالجة المسبقة للبيانات أو استخدام أساليب تحليل السلاسل الزمنية التي تأخذ في الاعتبار التغيرات الموسمية يمكن أن يساعد على تقليل تأثير التحيز الفصلي.



تحيز استخراج البيانات

يحدث هذا التحيز عندما تقوم أنظمة الذكاء الاصطناعي أو من يقومون بتطويرها بتحليل مجموعات بيانات ضخمة، مع التركيز فقط على الأنماط أو النتائج التي تظهر أهمية إحصائية، دون مراعاة السياق الكامل أو احتمالية العشوائية. يؤدي ذلك إلى تدريب النماذج على علاقات زائفة أو مضللة، مما يضعف دقة تنبؤات الذكاء الاصطناعي وموثوقية مخرجاته.

يؤدي هذا التحيز إلى خلق صورة مضللة حول فاعلية النموذج، حيث يبدو أنه يحقق أداءً ممتازاً داخل بيانات التدريب، لكنه يفشل في التعميم على بيانات جديدة. وقد يؤدي ذلك إلى قرارات خاطئة في البيئات الواقعية، مما يضعف الثقة بأنظمة الذكاء الاصطناعي، خاصة عند استخدامها في مجالات حساسة تتطلب دقة عالية وتعميماً فعالاً.



التأثير

يقوم الباحثون بتجربة فرضيات مختلفة حول مدى فاعلية الميزات المختلفة في التنبؤ بمشتريات العملاء، من خلال تحليل البيانات عدة مرات، وتجربة مجموعات مختلفة من المتغيرات حتى يتوصلوا إلى نتيجة ذات دلالة إحصائية. وبالرغم من ذلك، فإنهم يرفعون تقاريراً بالنتائج المهمة ويتجاهلون المحاولات الفاشلة. وهذا الرفع الانتقائي للتقارير قد يؤدي إلى توهم وجود أنماط ذات معنى، مما يؤدي إلى استنتاجات خاطئة.



المثال

تحديد الفرضيات مسبقاً، واستخدام تقنيات تحقق مثل تقسيم البيانات أو التحقق المتبادل، وتطبيق تصحيحات للمقارنات المتعددة، مع الإبلاغ بشفافية عن جميع التحليلات، بما في ذلك النتائج غير الناجحة، لضمان مصداقية نماذج الذكاء الاصطناعي.



التخفيف

تحيز الرقابة الأخلاقية

يحدث هذا التحيز عندما تفتقر أنظمة الذكاء الاصطناعي إلى آليات رقابة أو إشراف فعالة لضمان التزامها بالسلوك الأخلاقي، والقوانين، واللوائح، أو الإرشادات الأخلاقية ذات الصلة. يؤدي هذا النقص إلى مخاطر في اتخاذ قرارات غير مسؤولة أو غير متسقة مع المبادئ الإنسانية.

يؤثر هذا التحيز في سلامة القرارات الصادرة عن أنظمة الذكاء الاصطناعي، إذ يسمح لها بالعمل دون ضوابط أو مراجعات أخلاقية كافية. يمكن أن يؤدي ذلك إلى سلوكيات ضارة أو غير مسؤولة تنتهك المعايير الإنسانية، وتسبب أضراراً اجتماعية أو قانونية أو مؤسسية، وتؤثر في الثقة والقبول العام للتقنية.



التأثير

نموذج ذكاء اصطناعي يُستخدم في تصنيف السجلات الطبية تم نشره دون مراجعة أخلاقية كافية لضمان حماية خصوصية المرضى. لم يُؤخذ في الحسبان أن النموذج قد يستنتج معلومات حساسة بشكل غير مباشر، مثل الحالة النفسية أو التاريخ الصحي، دون موافقة صريحة، إذ إن غياب الرقابة الأخلاقية أدى إلى انتهاك خصوصية الأفراد، وزعزع الثقة في استخدام الذكاء الاصطناعي في القطاع الصحي.



المثال

تطبيق أطر حوكمة قوية تتضمن آليات رقابة تمكن من مراقبة وضمان أن أنظمة الذكاء الاصطناعي تلتزم بالمبادئ الأخلاقية طوال عملية تطويرها، ونشرها، وتشغيلها. هذا يتضمن وضع سياسات وإرشادات واضحة لتطوير الذكاء الاصطناعي بشكل أخلاقي، وضمان وجود آليات رقابة مثل لجان الاستعراض الأخلاقي لتقييم مشاريع وقرارات الذكاء الاصطناعي من منظور أخلاقي. ومن المهم أيضاً تعزيز التعليم والتدريب المستمر حول مبادئ الذكاء الاصطناعي الأخلاقي للمطورين والمستخدمين واتخاذ القرارات المعنية بمشاريع الذكاء الاصطناعي.



التخفيف

تحيز تسرب البيانات

يشير تحيز تسرب البيانات إلى الحالة التي يتم فيها تضمين معلومات من بيانات الاختبار أو الإنتاج ضمن بيانات التدريب، مما يؤدي إلى إنشاء نموذج يبدو عالي الدقة أثناء التقييم، لكنه يفشل عند تطبيقه على بيانات جديدة. هذا التحيز يضر بقدرة النموذج على التعميم ويؤدي إلى نتائج غير موثوقة في البيئات الواقعية.

يؤدي هذا التحيز إلى إضعاف موثوقية النموذج عند التعامل مع سيناريوهات جديدة، إذ يمنحه ميزة غير عادلة أثناء التدريب قد لا تعكس أدائه الفعلي لاحقاً. ينتج عن ذلك فجوة بين الأداء المتوقع والأداء الواقعي، ما قد يتسبب في اتخاذ قرارات غير دقيقة في التطبيقات الحرجة.



التأثير

نموذج مصمم للتنبؤ بإعادة الدخول إلى المستشفى: إذا كانت بيانات المريض المستخدمة للتدريب تتضمن معلومات إعادة الدخول كأحدى خصائص الإدخال، فقد يتعلم النموذج ربط تفاصيل محددة في البيانات بالنتيجة مباشرة، وسيعمل النموذج بشكل جيد للغاية أثناء التحقق من الصحة لأنه قد تم إعطاؤه الإجابات مسبقاً. ومع ذلك، لن يستخدم هذا الأداء في المواقف الواقعية حيث لا تتوفر بيانات إعادة الدخول بشكل مسبق، مما يؤدي إلى انخفاض كبير في فاعلية أداء النموذج.



المثال

الالتزام بتقسيم مجموعة الاختبار ومجموعة التدريب بشكل مناسب، مما يضمن تطبيق المعالجة المسبقة على بيانات التدريب فقط. ويجب تجنب اختيار الميزات التي قد تؤدي إلى تسرب المعلومات المتغيرة المستهدفة. كما يساعد تنفيذ خطوات معالجة بيانات منفصلة للمعالجة المسبقة والتدريب على الحفاظ على سلامة البيانات، مع خطوات المعالجة المسبقة مثل القياس بناءً على بيانات التدريب فقط. كذلك، تعد إجراءات التحقق الصارمة ضرورية لمعالجة تسرب البيانات، مما يضمن التقييم الدقيق لأداء النموذج.



التخفيف

تحيز لعبة الاستبدال

يشير هذا التحيز إلى ميل نماذج الذكاء الاصطناعي، خاصة نماذج اللغة، إلى استبدال الكلمات أو المفاهيم اعتماداً على الارتباطات الإحصائية بدلاً من الفهم العميق للسياق. يؤدي ذلك إلى مخرجات قد تبدو منطقية ظاهرياً لكنها تفتقر إلى الدقة أو الفهم الحقيقي للمضمون.

يؤثر هذا التحيز في دقة المخرجات، حيث قد تولد نماذج الذكاء الاصطناعي استجابات تبدو منطقية شكلياً لكنها تفتقر إلى المعنى الحقيقي. يمكن أن يؤدي ذلك إلى إساءة الفهم أو الاستنتاج الخاطئ، خاصة في المجالات التي تتطلب تفسيراً دقيقاً أو تحليلاً سياقياً للبيانات.



التأثير

نموذج ذكاء اصطناعي مدرب على بيانات طبية لتقديم إجابات حول الصحة، عندما يسأل المستخدم النموذج: "هل الألم في الصدر علامة على نوبة قلبية؟" قد يجيب النموذج: "نعم، الألم في الصدر مرتبط بالنوبات القلبية." هذه الإجابة تبدو منطقية لكنها تفتقر إلى العمق، لأن النموذج يكرر نمطاً شائعاً في بياناته التدريبية دون فهم السياق الطبي. في الواقع، يمكن أن يكون ألم الصدر علامة على عدد من الحالات الأخرى. إذا كان النموذج لا يفهم الفروقات فقد لا يقترح مجموعة من الاحتمالات الأخرى.



المثال

الالتزام بتعريفات دقيقة للمتغيرات والمفاهيم المستخدمة، واختيار تصاميم دراسية متكاملة، إلى جانب مراجعة الأقران المتخصصين وتحليل البيانات بدقة. كما أن الشفافية في توثيق التعديلات والتواصل الواضح بشأن التغييرات المنهجية تساهم في تعزيز موثوقية النتائج وجودة المخرجات.



التخفيف

تحيز لغوي

يشير إلى التحيز الذي ينبع من التفاصيل والاختلافات في اللغة البشرية التي يتم تدريب أنظمة الذكاء الاصطناعي عليها. يظهر هذا التحيز عندما تتبنى أنظمة معالجة اللغات الطبيعية النماذج النمطية والافتراضات الثقافية أو التمثيلات غير المتساوية الموجودة في بيانات التدريب الخاصة بها. نظراً إلى أن اللغة تعكس بعمق المواقف والتحيزات الاجتماعية، فإن نماذج الذكاء الاصطناعي قد تتعلم هذه الأنماط بشكل غير مقصود، مما يؤثر في فهمها وتوليدها للنصوص.

يؤدي هذا التحيز إلى نتائج غير عادلة عند تحليل النصوص أو التعامل معها، حيث تميل الأنظمة إلى تكريس الفروقات الثقافية والتمثيلات غير المتوازنة الموجودة في بيانات التدريب. وقد يظهر الأثر بشكل بارز في التفاعل مع المستخدمين أو عند تقديم توصيات، مما قد يعزز الصور النمطية أو يقصي أصواتاً مهمة.



نموذج ذكاء اصطناعي لتحليل مراجعات العملاء تم تدريبه على نصوص مكتوبة بلغة رسمية وفصحى فقط. عند استخدامه لتحليل تعليقات مكتوبة بلهجات عامية أو بأسلوب غير رسمي، يفشل النموذج في فهم المعنى بدقة أو يصنف المراجعة بشكل خاطئ. هذا التحيز اللغوي يؤدي إلى استجابات غير دقيقة ويهمش أنماط التعبير الشائعة بين فئات واسعة من المستخدمين.



تنوع مجموعات البيانات التدريبية لتشمل مجموعة واسعة من التعبيرات اللغوية من مختلف الثقافات والأجناس والفئات الاجتماعية، مما يضمن تعرض نظام الذكاء الاصطناعي لتمثيل متوازن لاستخدام اللغة. بالإضافة إلى ذلك، فإن استخدام تقنيات إزالة التحيز أثناء التدريب النموذجي، مثل تعديل تضمينات الكلمات لتحديد التحيز أو تطبيق قيود الإنصاف، يمكن أن يساعد على تقليل استمرار الصور النمطية الضارة.



تحيز مرشح الإحالة

يحدث تحيز مرشح الإحالة عندما تُبنى بيانات التدريب في تطبيقات الذكاء الاصطناعي على حالات تم اختيارها مسبقاً من خلال عملية إحالة لا تعكس بالضرورة خصائص العينة الأوسع، مما يؤدي إلى تدريب النموذج على بيانات غير ممثلة ويرفع من احتمالية إصدار تنبؤات أو نتائج متحيزة أو غير دقيقة.

يُضعف هذا التحيز من قابلية تعميم النموذج، إذ يُدرب على عينات غير ممثلة، غالباً نتيجة اختيار بيانات عبر عملية إحالة، يترتب على ذلك أداء محدود خارج هذه العينة، ما قد يؤدي إلى نتائج خاطئة أو غير دقيقة في التنبؤات أو التشخيصات، خاصة في تطبيقات الرعاية الصحية أو الموارد البشرية.



التأثير

نظام الذكاء الاصطناعي الذي تم تطويره للتنبؤ بمخاطر الإصابة بأمراض القلب على بيانات المرضى الذين تمت إحالتهم إلى أطباء القلب بسبب أعراض أو عوامل خطر محددة، فإذا كان نمط الإحالة يعكس تحيزات -على سبيل المثال: تقل احتمالية إحالة المرضى الأصغر سناً-، فقد يقلل نموذج الذكاء الاصطناعي من خطر الإصابة بالمرض لدى السكان الأصغر سناً، مما يؤثر في جودة الرعاية والتدابير الوقائية المقدمة لهم.



المثال

تضمين مجموعة بيانات متنوعة وتمثيلية تغطي النطاق الكامل للسكان، وليس فقط أولئك الذين تمت إحالتهم لحالة أو علاج معين. وقد يتضمن ذلك جمع البيانات من مجموعة متنوعة من البيئات، بما في ذلك مراكز الرعاية الأولية والصحة المجتمعية، لالتقاط نطاق أوسع من تجارب المرضى. بالإضافة إلى ذلك، فإن تطبيق الأساليب الإحصائية لضبط عملية الإحالة أو استخدام تقنيات مثل توليد البيانات الاصطناعية لتمثيل المجموعات الممثلة تمثيلاً ناقصاً في بيانات التدريب يمكن أن يساعد على تقليل تأثير تحيز مرشح الإحالة.



التخفيف

تحيز معدل القبول

يشير إلى أي حالة يظهر فيها نظام الذكاء الاصطناعي تحيزاً في القبولات المتعلقة بالجامعات، أو الوظائف، أو أي سيناريو يتم فيه قبول أو رفض الأفراد بناءً على معايير معينة. يحدث ذلك عندما تكون نتائج نظام الذكاء الاصطناعي ماثلة بشكل منهجي، مفضلة أو محقة بحق مجموعات معينة من الأفراد بناءً على الجنس، أو العرق، أو الوضع الاجتماعي والاقتصادي.

يعزز هذا التحيز التفاوتات الاجتماعية، عندما يُظهر نظام الذكاء الاصطناعي تحيزاً في قرارات القبول للجامعات أو الوظائف. كما يؤدي ذلك إلى ترسيخ الامتيازات المسبقة أو حرمان مجموعات معينة من فرص عادلة، مما يضر بمبدأ المساواة ويُنتج مخبرات قد تفتقر إلى العدالة أو الدقة.



التأثير

نموذج ذكاء اصطناعي يُستخدم في تصفية طلبات التوظيف يعتمد على بيانات تاريخية من موظفين تم توظيفهم سابقاً من جامعات معينة أو شركات مرموقة. يؤدي ذلك إلى تفضيل المتقدمين من خلفيات مماثلة، وتجاهل مرشحين مؤهلين من جامعات أقل شهرة أو من خلفيات اجتماعية واقتصادية متواضعة، مما يُنتج قرارات قبول غير عادلة تُكرس الفجوة في فرص التوظيف.



المثال

اختيار بيانات التدريب المتنوعة والتي تعكس سكان العالم الحقيقي، مما يضمن أن تكون قرارات نظام الذكاء الاصطناعي عادلة وشاملة. يعد الرصد والتحليل المنتظم لنتائج النظام يعد أمراً بالغ الأهمية للكشف عن أي تحيزات، مع إجراء التعديلات حسب الضرورة. تساعد هذه العملية على ضمان أن القرارات المتعلقة بالقبول لا تحابي أو تحرم أي مجموعة بشكل غير عادل.



التخفيف

تحيز معرفي

ينشأ التحيز المعرفي عندما يعتمد الدماغ البشري على أنماط تفكير مبسطة لتسريع معالجة المعلومات عبر فلترة التجارب والتفضيلات الشخصية. تساعد هذه الفلترة على تسريع معالجة كميات كبيرة من المعلومات وتحديد الأولويات، إلا أن هذه الآلية، رغم فعاليتها، قد تؤدي إلى أخطاء في التفسير واتخاذ القرار.

ينعكس هذا التحيز في طريقة تصميم الأنظمة وتفسير مخرجاتها، نتيجة تدخل التصورات الذاتية للمطورين أو محللي البيانات. على الرغم من كونها جزءاً من الآليات الإدراكية البشرية، إلا أن التحيزات المعرفية قد تشوه خوارزميات الذكاء الاصطناعي وتؤدي إلى استنتاجات متحيزة أو غير دقيقة.



التأثير

عندما يكون هناك نظام ذكاء اصطناعي للموافقة على القروض، فإذا كان لدى المطورين اعتقاد باطني بأن مجموعات سكانية معينة أقل جدارة ائتمانية، فإن هذا التحيز يمكن أن يؤثر على اختيار بيانات التدريب أو تصميم الخوارزمية، مما يؤدي إلى نظام يميز بشكل غير عادل ضد المتقدمين من تلك المجموعات حتى إذا لم تكن هناك بيانات تمييزية.



المثال

تثقيف مطوري الذكاء الاصطناعي وأصحاب المصلحة حول التحيزات المعرفية وتأثيرها المحتمل في أنظمة الذكاء الاصطناعي أمر بالغ الأهمية. يمكن أن يساعد الوعي البشري على التعرف على تحيزاتهم والتخفيف منها خلال دورة تطوير الذكاء الاصطناعي. علاوة على ذلك، فإن دمج فرق متنوعة في أبحاث الذكاء الاصطناعي قد يساعد على تخفيف التحيزات الفردية من خلال تقديم مجموعة متنوعة من وجهات النظر لعملية صنع القرار، بدءاً من اختيار البيانات وحتى تصميم الخوارزميات وتفسير النتائج.



التخفيف

تحيز نتائج التقرير

يشير تحيز نتائج التقرير إلى الميل إلى عرض نتائج تطبيقات الذكاء الاصطناعي التي تُظهر أداءً مرتفعاً أو تأثيراً إيجابياً، مع إغفال النتائج السلبية أو غير الحاسمة. هذا التحيز يؤدي إلى رسم صورة غير دقيقة لقدرات النظام، ويؤثر في تقييم الباحثين وصناع القرار، كما يعزز التوقعات بشكل مبالغ فيه لدى المستخدمين والجمهور.

يؤدي هذا التحيز إلى تقديم صورة غير مكتملة عن فاعلية أنظمة الذكاء الاصطناعي، نتيجة التركيز على النتائج الإيجابية وتجاهل ما دونها. يؤثر ذلك في قرارات السياسات العامة وتمويل الأبحاث، ويغذي تصورات مضللة لدى المستخدمين وصناع القرار حول إمكانيات التقنية وحدودها.



التأثير

نظام ذكاء اصطناعي للتنبؤ بتوجهات سوق الأسهم، ويقوم المطورون بنشر الحالات التي كانت توقعاته فيها دقيقة فقط، مع إخفاء أو تجاهل التنبؤات غير الصحيحة، مما يؤدي إلى انطباع زائف حول فاعلية النموذج، ويُضلل المستثمرين أو متخذي القرار بشأن دقته وموثوقيته في الواقع العملي.



المثال

الإفصاح عن جميع التفاصيل ذات الصلة بالبحث، بما في ذلك النتائج السلبية وإجراءات تنظيف البيانات ومعايير اختيار النموذج. كذلك، تحديد أسئلة البحث والفرضيات وطرق جمع البيانات وخطة التحليل بوضوح قبل جمع البيانات. وهذا يقلل من خطر الإبلاغ الانتقائي واستخراج البيانات. كما يمكن، تسليط الضوء على قيود الدراسة والتأثير المحتمل للتحيزات واتباع المبادئ التوجيهية الأخلاقية لبحوث الذكاء الاصطناعي، مثل تلك التي تقدمها منظمات مثل ACM أو IEEE. بالإضافة إلى التأكد من تطوير أنظمة الذكاء الاصطناعي ونشرها بطريقة عادلة وغير متحيزة.



التخفيف

تحيز نحو الشعبية

يحدث ذلك عندما تقوم أنظمة الذكاء الاصطناعي بتسليط الضوء على العناصر الشائعة أو التوصية بها بشكل غير متناسب مع العناصر الأخرى، مما يؤدي إلى دورة تكتسب فيها هذه العناصر مزيداً من الاهتمام والشعبية، بينما تظل العناصر الأقل شعبية مجهولة. وهذا يمكن أن يؤدي إلى تحريف تمثيل الاختيارات المتاحة للمستخدمين، مما يحد من تنوع المحتوى أو المنتجات التي يتعرضون لها.

يساهم هذا التحيز في تضيق نطاق الخيارات المعروضة للمستخدمين، عبر تركيز أنظمة الذكاء الاصطناعي على العناصر الشائعة بشكل مفرط. يؤدي ذلك إلى نشوء "عرف الصدى" الرقمية، حيث يتم تكرار المحتوى ذاته دون تنوع كافٍ، مما يضعف تجربة المستخدم ويحد من الانفتاح على خيارات بديلة أو غير مألوفة.



التأثير

نموذج ذكاء اصطناعي يُستخدم في متجر إلكتروني للتوصية بالمنتجات يميل إلى اقتراح السلع الأكثر مبيعاً بشكل متكرر. مع مرور الوقت، تصبح هذه المنتجات أكثر ظهوراً وتزداد مبيعاتها فقط بسبب تعزيز النظام لها، بينما تُهمَّش منتجات أخرى قد تكون أكثر جودة أو ملاءمة لاحتياجات المستخدم. هذا التحيز يقلل من تنوع الخيارات ويحد من فرص المنتجات الجديدة أو المتخصصة في الظهور.



المثال

تعديل خوارزميات الذكاء الاصطناعي لإبراز العناصر الأقل شهرة جنباً إلى جنب مع العناصر الشائعة، مما يضمن تمثيل نطاق أوسع وأكثر تنوعاً من المحتوى. يمكن أن تساعد التوصيات التي تركز على التنوع وآليات التغذية الراجعة على إعادة تقييم وتعديل مستوى العرض الممنوح للعناصر بشكل دوري، مما يعزز توازن الرؤية عبر جميع العناصر.



التخفيف

تحيز نحو المركز

يشير إلى الظاهرة التي تميل فيها نتائج النماذج الخوارزمية إلى التقارب حول اتجاه سائد، أو وجهة نظر، أو نقطة مركزية. يمكن أن يظهر هذا التحيز بطرق مختلفة، من تمثيل زائد لآراء الأغلبية في مجموعة البيانات إلى خوارزميات تفضل بشكل غير متناسب النتائج التي تتماشى مع الاتجاهات الأكثر شيوعاً أو انتشاراً.

يضعف هذا التحيز قدرة النماذج على استيعاب التنوع المعرفي والاجتماعي، إذ تميل الخوارزميات إلى الاتجاهات المركزية السائدة، متجاهلة الأصوات الهامشية أو غير النمطية. هذا يحد من فاعلية الذكاء الاصطناعي في التقاط الأنماط الدقيقة ويقلل من إمكانياته في معالجة الظواهر المعقدة.



التأثير

نموذج ذكاء اصطناعي لتحليل الرأي العام تم تدريبه على تعليقات من منصات تواصل اجتماعي عالمية كبرى ناطقة بالإنجليزية، مما أدى إلى تعزيز وجهات النظر الشائعة في المجتمعات الغربية وتجاهل الآراء أو التعبيرات الأقل شيوعاً أو القادمة من ثقافات أخرى. هذا التحيز يقلل من قدرة النموذج على فهم التنوع الثقافي واللغوي ويؤدي إلى تحليلات غير متوازنة عند استخدامه عالمياً.



المثال

جمع البيانات المتنوعة مما يضمن التمثيل من مجموعة واسعة من التركيبة السكانية لتجنب المبالغة في التركيز على خصائص المجموعة المركزية أو المهيمنة. يعد تنفيذ تقنيات اكتشاف التحيز وتصحيحه أمراً ضرورياً، بما في ذلك المراقبة المنتظمة لأداء النموذج والشفافية الخوارزمية لتحديد التحيزات ومعالجتها.



التخفيف

تحيز نحو المستقبلات

يحدث التحيز نحو المستقبلات عندما تعطي النماذج أو الخوارزميات أولوية مفرطة للبيانات الجديدة أو النادرة أثناء التدريب أو اتخاذ القرار، على حساب البيانات الأكثر شيوعاً أو صلة بالسياق. يؤدي هذا التحيز إلى تضخيم أهمية المعلومات الحديثة، مما قد يؤثر سلباً في قدرة النموذج على التعميم ويقلل من دقته في المواقف الواقعية.

يفاقم هذا التحيز من تضخيم أهمية البيانات الحديثة أو النادرة على حساب السياق الطويل الأمد. قد يؤدي إلى نتائج مضللة، لا سيما في التحليلات التنبؤية، حيث تُمنح الأحداث الحديثة وزناً غير مبرر، مما يُضعف دقة النموذج واستقراره في البيئات الواقعية.



التأثير

عندما يستخدم نموذج لتعلم الآلة للتنبؤ بسوق الأوراق المالية، وكان هذا النموذج يحمل تحيزاً نحو المستقبلات، فقد يبالغ في تقدير الحالات الشاذة الأخيرة في السوق باعتبارها تنبؤاً بالتوجهات المستقبلية، متجاهلاً الأنماط السابقة الأكثر اتساقاً التي توفر أساساً أفضل للتنبؤات. وقد يؤدي هذا إلى توقعات غير دقيقة واستراتيجيات استثمار مضللة استناداً إلى افتراض أن هذه الحالات الشاذة سوف تتكرر بشكل منتظم.



المثال

تصميم أنظمة الذكاء الاصطناعي لتقييم ووزن نقاط البيانات الجديدة أو النادرة في سياق أهميتها الشاملة ومساهمتها في أهداف النموذج. واستخدام تقنيات مثل ضبط معدل تعلم النموذج للمعلومات الجديدة، واستخدام أساليب التنظيم لمنع الإفراط في البيانات الجديدة، والتحقق من صحة أداء النموذج عبر مزيج متوازن من السيناريوهات الشائعة والنادرة يمكن أن تساعد على ضمان عدم تأثير الحداثة بشكل غير ضروري في تنبؤات النموذج.



التخفيف

تحيز نحو النتائج الإيجابية

يشير هذا التحيز إلى انحياز في نظام الذكاء الاصطناعي يؤدي إلى تضخيم عرض النتائج الإيجابية أو المؤكدة مقارنةً بالنتائج السلبية أو غير الحاسمة. يحدث ذلك عندما يطور النظام آلياته الداخلية لترجيح أو إبراز المخرجات التي تظهر نجاحاً، مع تقليل أو تجاهل الحالات التي تظهر فشلاً أو ضعفاً في الأداء، مما يؤدي إلى تقديم صورة مضللة عن فاعلية النموذج ويؤثر في دقة التقييم واتخاذ القرار.

يشوه هذا التحيز صورة الأداء الفعلي للنماذج، نتيجة الإفراط في عرض النتائج الإيجابية وإغفال السلبية أو غير الحاسمة. هذا يخل بسلامة القرارات المستندة إلى مخرجات الذكاء الاصطناعي، ويؤثر سلباً في الأبحاث العلمية، وتقييمات السياسات، والمعايير الأخلاقية في نشر النتائج.



التأثير

نموذج ذكاء اصطناعي للتشخيص الطبي يتم تقييمه باستخدام نتائج حالات نجح في تشخيصها فقط، بينما يتم تجاهل الحالات التي فشل فيها في تحديد المرض أو خطأ في التصنيف. هذا التحيز يؤدي إلى تضخيم دقة النموذج الظاهرة، ويخفي محدودياته، مما قد يؤثر سلباً في قرارات الأطباء وثقة المرضى، ويؤدي إلى اعتماد غير مدروس على النظام في مواقف حرجية.



المثال

الالتزام بالنشر المفتوح لجميع نتائج النماذج، سواء كانت إيجابية أو سلبية، يضمن الشفافية ويتيح تقييماً شاملاً للأبحاث. ويساعد هذا النهج على تقليل التحيز نحو النتائج الإيجابية ويشجع على تقديم صورة أكثر دقة وشمولية عن أداء النماذج. كما أن تحسين معايير النشر يتطلب التركيز على جودة البيانات المستخدمة، وتصميم النماذج، واختبارات التقييم بدلاً من السعي وراء النتائج المثيرة فقط. ويعزز هذا من موثوقية الدراسات ويساعد على بناء ثقة أكبر في الأبحاث المنشورة.



التخفيف

تحيز وقت الاستجابة

يحدث هذا التحيز عندما يتم تقييم نظام الذكاء الاصطناعي على أنه أكثر فاعلية مما هو عليه فعلياً، بسبب اكتشاف مبكر لحالة أو ظاهرة ما. فقد يبدو أن النظام يُحسن النتائج أو يمدد فترة النجاح، في حين أن التحسن الظاهري يعود فقط إلى التوقيت المبكر للاكتشاف، وليس لتأثير حقيقي في مسار الحالة أو مخرجاتها، وقد يؤدي هذا إلى تقدير مضلل لأداء النظام.

يخلق هذا التحيز انطباعاً مضللاً عن فاعلية النماذج، حيث تُعزى تحسينات النتائج إلى الاكتشاف المبكر، في حين أن التوقيت هو العامل الحاسم وليس تدخل النموذج بحد ذاته. يؤدي ذلك إلى توصيات قد تفتقر للدقة أو تُبالغ في فاعلية التدخلات المقترحة.



التأثير

نموذج ذكاء اصطناعي مخصص لاكتشاف الأورام يُظهر تحسناً في نتائج البقاء على قيد الحياة عند استخدامه، لكن هذا التحسن لا يعود إلى فاعلية النموذج في تغيير مسار المرض، بل فقط لأنه يكتشف الورم في وقت أبكر من المعتاد. التقييم المبني على التوقيت المبكر يعطي انطباعاً مضللاً بفاعلية النظام، بينما النتائج الفعلية للعلاج تبقى دون تغيير.



المثال

زيادة الوعي بين الباحثين والممارسين حول التحيز الناتج عن وقت الاستجابة وكيفية التخفيف منه يمكن أن يساهم في تحسين تصميم الدراسات وتحليل النتائج. تحليل الحساسية يقوم بتقييم مدى تأثير التحيز الناتج عن وقت الاستجابة في النتائج المتوقعة باستخدام نماذج إحصائية لمحاكاة سيناريوهات مختلفة وتقدير التأثيرات المحتملة.



التخفيف

تحيز وقت البدء

يشير هذا التحيز إلى انحراف في أداء نظام الذكاء الاصطناعي ناتج عن توقّيت بدء جمع البيانات أو تدريب النموذج على ظاهرة معينة خلال فترة زمنية محددة. إذا تم بناء النظام استناداً إلى بيانات من فترة غير ممثلة (مثل وقت أزمة أو ذروة موسمية)، فقد يتعلم النموذج أنماطاً مؤقتة ويظهر سلوكاً منحازاً عند تطبيقه على فترات مختلفة، مما يؤدي إلى تنبؤات أو قرارات غير دقيقة.

يؤثر هذا التحيز في دقة التحليلات الزمنية، حيث تؤدي بداية الدراسة في لحظة غير تمثيلية إلى نتائج منحرفة لا تعكس واقع الظاهرة المدروسة. كما أنه يؤثر في القرارات السياسية أو التشغيلية المعتمدة على تلك البيانات، ويقلل من موثوقية التنبؤات المستخلصة منها.



التأثير

نموذج ذكاء اصطناعي لتحليل أداء الطلاب تم تدريبه على بيانات جمعت في بداية العام الدراسي فقط، حيث يكون الطلاب في مرحلة التكيف والتأقلم الذهني أعلى. عند تطبيق هذا النموذج لاحقاً خلال الفصل، قد يُظهر سلوكاً منحازاً ويقلل من دقة التقييم، لأنه تعلم أنماطاً مرتبطة بزمان محدد لا تمثل الأداء العام للطلاب على مدار العام، مما يؤدي إلى قرارات غير دقيقة حول احتياجاتهم التعليمية.



المثال

ضبط متغيرات البداية للمشاركين مثل العمر والحالة الصحية وتوحيد فترات البداية لمجموعات المقارنة قدر الإمكان. بالإضافة إلى استخدام تصاميم دراسية تسمح بالمقارنات الزمنية أو التدخلات العشوائية. كما يمكن تحليل البيانات بدقة لتصحيح أي تأثيرات محتملة ناجمة عن اختلافات في البداية وإجراء تقييمات مستمرة للتأثيرات المحتملة للتأكد من دقة النتائج. باستخدام هذه الإجراءات، يمكن للدراسات أن تحقق نتائج أكثر دقة وموضوعية، متجنباً التحيز الناتج عن فترات البداية المختلفة بين المجموعات.



التخفيف

- Zhang, J. et al. Robustness of Fairness in AI: A Study of Classification Problems (2022). Available at: <https://arxiv.org/pdf/2201.11706.pdf>
- Binns, R. On the Definition of AI Bias and Its Ethical Implications. Journal of AI and Society 34, 99-108 (2019). Available at: <https://link.springer.com/content/pdf/10.1007/s11023-019-09506-6.pdf>
- Catalog of Bias. Biases (n.d.). Available at: <https://catalogofbias.org/biases/>
- Cowgill, B. et al. Bias in Machine Learning: A Survey of Current Research. arXiv (2020). Available at: <https://arxiv.org/abs/2011.11486>
- Taylor, C. AI Ethics: Key Concepts and Considerations. In: The Oxford Handbook of Ethics of AI. Available at: <https://www.oxfordreference.com/display/10.1093/acref/9780191843273.001.0001/acref-9780191843273-e-143>
- IBM. AI Bias: Understanding the Risks(2023). Available at: <https://www.ibm.com/topics/ai-bias>
- IBM. Explaining AI Bias with Real-World Examples(2020). Available at: <https://www.ibm.com/downloads/cas/3KY8VLZY>
- Adams, R. & Wilson, M. Ethical and Policy Issues in AI In: Handbook of Ethical AI 7-21 (2023). Available at: https://link.springer.com/chapter/10.1007/978-3-031-39355-6_7
- Zhang, H. et al. AI and Its Ethical Dilemmas: Understanding the Emerging Trends. arXiv (2023). Available at: <https://arxiv.org/abs/2308.01118>
- IBM. Shedding Light on AI Bias with Real-World Examples (2023). Available at: <https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/>
- IBM. Understanding Bias in AI Systems (2023). Available at: <https://www.ibm.com/topics/ai-bias>
- Shen, X. et al. Ethical Considerations in AI: A Systematic Review. NCBI PMC (2021). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9196683/>



للاستفسارات:

Studies@sdaia.gov.sa

نأمل مشاركتنا رأيك حول هذا
الدليل من خلال مسح الرمز
لتعبئة الاستبيان.



تاريخ الوثيقة | مايو 2026
نوع الوثيقة | أدلة إرشادية
تصنيف الوثيقة | عام
رقم الوثيقة | SDAIA-P103A
رقم الإصدار | 1



SDAIA.GOV.SA



SDAIA_SA



SDAIA.SAUDI



SDAIA-KSA