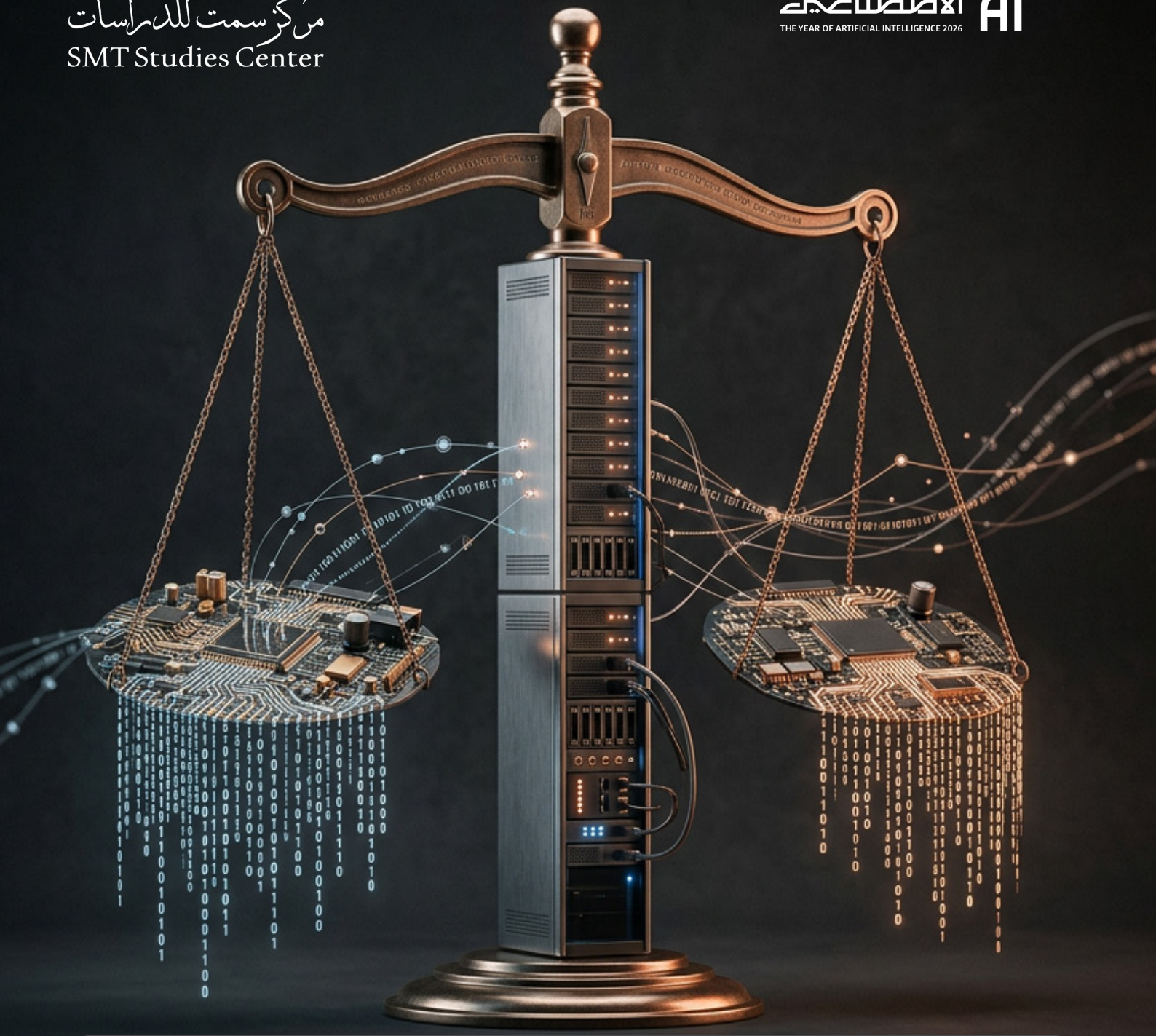
مركز سمت للدراسات
SMT Studies Center

عام الذكاء الاصطناعي
THE YEAR OF ARTIFICIAL INTELLIGENCE 2026



فخ "التحـــــيز الرقمي"

يوليو 2026

قراءة خاصة 269

مقدمة

لقد تم تأطير الانتقال من الحكم البشـــري إلى اتخاذ القرار الخوارزمي في البداية علـــى أنه انتصار للحياد. فمن خلال إزالة الطبيعـــة "الفوضوية" و"الذاتية" للتحيز البشري، جادل المؤيدون بأن الخوارزميات ســـتوفر منصة محايدة لتقييم الجدارة والمخاطـــر والإمكانـــات. ومع ذلـــك، فإن "حياديـــة" الخوارزمية هـــي مغالطة. فالخوارزميات ليست عوامل مستقلة؛ إنها انعكاسات للبيانات التي تستهلكها.

المشكلة الأساســـية التي تتناولها هذه الورقة هي "فخ التحيز الخوارزمي". يحدث هذا الفخ عندما تُســـتخدم البيانات التاريخية، التي غالبًا ما تكون ســـجلاً للتمييز السابق، لتدريب النماذج التي تتنبأ بالنتائج المستقبلية. عندما تتنبأ خوارزمية بمن هو الأرجح أن ينجح في وظيفة أو بمن هو الأرجح أن يعاود الإجرام، فإنها غالبًا ما تحدد ببساطة أنماط الامتياز السابق أو المبالغة في المراقبة.

المنهجية

للتحقيــق في هذه الظاهرة، تســتخدم هذه الدراسة تحليل الوثائق النوعي "QDA". تحليل الوثائق النوعي هو إجراء منهجي لمراجعة أو تقييم الوثائق، المطبوعــة والإلكترونية على حد ســواء. هذه الطريقة مناسبة بشكل خاص لهــذا البحث لأنها تســمح بتفســير كيفية تعريف "التحيــز" ومعالجته عبــر القطاعات المختلفة.

تتضمن عمليــة تحليــل الوثائق النوعي لهذه الورقة ما يلي:

اختيار الوثائق: لقد قمنا بتجميع مجموعة من الوثائق بمــا في ذلك المجلات المحكمة حول أخلاقيات الذكاء الاصطناعي، والأوراق البيضاء من شــركات التكنولوجيا الكبرى "مثل Google، IBM"، وتقارير السياسات من منظمات الحقوق المدنيــة "مثــل الاتحاد الأميركــي للحريات المدنية ACLU".

الترميز الموضوعي: يتــم تحليل الوثائق بحثًا عــن الموضوعات المتكررة مثل "الشــفافية"، و"دقة التنبؤ"، و"التعويض التاريخي"، و"العدالة الرياضية".

التقييم السياقي: بدلاً من مجرد عــد الكلمات الرئيسية، يركز التحليل على نية الوثائق، سواء كانت تتعامل مع التحيز كـ "خلل" تقني يجب إصلاحه أو "ميزة" نظامية للبيانات.

التثليث: مــن خلال مقارنــة الوثائق الفنية بالنقــد المتعلق بالعدالــة الاجتماعية، يحدد البحــث "الفجــوة" بيــن ما يمكــن أن تحله الخوارزمية وما يتطلبه المجتمع.

أبعاد الظلم الخوارزمي

تنــدرج الأدبيــات الحاليــة حــول التحيــز الخوارزمي عمومًا ضمن ثلاث فئات.. التقنية، والقانونيــة، والاجتماعية. يكشــف تحليلنا لهــذه الوثائق عــن توتر كبير بيــن كيفية تعريف المهندســين لــ "الإنصاف" وكيفية تعريف علماء الاجتماع لـ "العدالة".

أسطورة حيادية البيانات

البيانات كمرآة: يجادل العديد من الباحثين بأن البيانات ليست "خامًا" بل "مُعدّة". إنها تعكس تحيزات البشر الذين جمعوها والمؤسسات التي أنتجتها.

المتغيرات الوكيلة: حتى عندما تتم إزالة السمات الحساسة مثل العرق أو الجنس، غالبًا ما تستخدم الخوارزميات "متغيرات وكيلة"، مثل الرموز البريديةٍ أو عادات التسوق، التي ترتبط ارتباطا وثيقا بالفئات المحمية، مما يؤدي إلى "التحديد العنصري" في العصر الرقمي.

حلقة التغذية الراجعة: في الشرطة التنبؤية، إذا أظهرت البيانات التاريخية المزيد من الاعتقالات في حي معين بسبب الإفراط في المراقبة، فإن الخوارزمية ستقترح إرسال المزيد من الشرطة إلى هناك، مما يخلق نبوءة تحقق ذاتها.

الإنصاف الفني مقابل الإنصاف الجوهري

التكافؤ الإحصائي: هذا هو الحل التقني الأكثر شيوعًا، حيث يضمن أن يكون "معدل النجاح" متساويًا بين المجموعات. ومع ذلك، يظهر تحليل الوثائق النوعي للأوراق التقنية أن هذا غالبًا ما يتجاهل الأسباب الكامنة وراء التفاوت.

مشكلة الصندوق الأسود: إن الافتقار إلى قابلية التفسير في نماذج التعلم العميق يجعل من الصعب على الأفراد الاعتراض على قرار ما. إذا تم رفض قرض بواسطة خوارزمية، فإن "السبب" غالبًا ما يكون مدفونًا في ملايين المعاملات الموزونة.

الأعمدة الثلاثة لفخ التحيز

من خلال تطبيق تحليل الوثائق النوعي على مجموعــة وثائقنا، تم تحديــد ثلاثة "فخاخ" رئيسية تمنع الخوارزميات من تحقيق العدالة الاجتماعية الحقيقية.

فخ قابلية النقل

- يحدث هذا عندما يتم تطبيق نموذج تم تطويره لسياق واحد "مثل أداة توظيف لشركة تقنية" على سياق مختلف "مثل وكالة حكومية" دون مراعاة الديناميكيات الاجتماعية المختلفة.
- غالبًا ما تعطي الوثائق الصادرة عن المطورين الأولوية "للقابلية للتوسع"، وهو ما يتعارض بشكل مباشر مع "السياق المحلي" المطلوب للعدالة الاجتماعية.

فخ الشكلية

- هذا هو الميل إلى الاعتقاد بأن المشكلات الاجتماعية مثل "الإنصاف" يمكن حلها بالكامل من خلال الصيغ الرياضية.
- يُظهر تحليلنا أن معظم الأطر التقنية تفشل في مراعاة "العدالة الإجرائية"، حق الشخص في أن يُسمع وأن يفهم المنطق وراء الحكم.

الفخ التاريخي

- الخوارزميات بطبيعتها تستند إلى الماضي. إنها تحسن "نسخة الماضي من النجاح".
- في التوظيف، إذا كان "أفضل أداء" للشركة في الماضي ينتمي بشكل أساسي إلى فئة ديموغرافية واحدة، فإن الخوارزمية ستستنتج رياضيًا أن تلك السمات الديموغرافية هي علامات للنجاح، وبالتالي تقوم بأتمتة "السقف الزجاجي".

نحو إطار للعدالة الاجتماعية

تشير نتائج تحليل الوثائق النوعي إلى أن "إصلاح" الخوارزمية ليس كافيا. يجب أن نتحرك نحو إطار للمساءلة الخوارزمية.

المبادئ المقترحة للعدالة الخوارزمية:

الإنسان في الحلقة: يجب أن تساعد الخوارزميات، لا أن تحل محل، الحكم البشري في المهام الحساسة، مما يسمح بمراعاة الظروف المخففة التي لا تستطيع البيانات التقاطها.

تقييمات الأثر الخوارزمي "AIAs": على غرار تقييمات الأثر البيئي، يجب أن يُطلب من المنظمات توثيق الأضرار الاجتماعية المحتملة للنموذج قبل نشره.

تصحيح البيانات: في الحالات التي يُعرف فيها أن البيانات التاريخية متحيزة "مثل السجلات الجنائية في بعض الولايات القضائية"، يقترح الباحثون "ترجيح" أو "بيانات اصطناعية" لمواجهة الاختلالات التاريخية بدلاً من مجرد عكسها.

حراس التوظيف الخوارزميون

غالبًا ما يتم تسويق الانتقال إلى التوظيف القائم على الذكاء الاصطناعي في وثائق الشركات كوسيلة "لإضفاء الطابع الديمقراطي" على التوظيف عن طريق إزالة "الحدس البشري". ومع ذلك، يكشف تحليلنا للأوراق البيضاء وتقارير التدقيق الخاصة بتقنيات الموارد البشرية أن هذه الأنظمة غالبًا ما تكرس الوضع الراهن.

تعريف "المواهب العليا" في الوثائق الفنية

هدف التحسين: يُظهر تحليل الوثائق النوعي لوثائق البائعين أن معظم خوارزميات التوظيف يتم تدريبها على "ملف تعريف النجاح" المستمد من الموظفين ذوي الأداء العالي الحاليين للشركة

إذا كانت القوى العاملة الحالية تفتقر إلى التنوع بسبب الإقصاء التاريخي، فإن الخوارزمية تتعلم أن خصائص المجموعة المهيمنة "مثل جامعات معينة، أو أنشطة خارج المنهج، أو حتى أنماط الكلام" هي المؤشرات الأساسية للنجاح.

تحليل السير الذاتية والتحيز الدلالي: يكشف تحليل نماذج معالجة اللغة الطبيعية "NLP" المستخدمة في التوظيف عن ميل لمعاقبة السير الذاتية التي تحتوي على لغة "جندرية". على سبيل المثال، أظهرت وثائق من عمليات تدقيق داخلية لشركات تقنية رفيعة المستوى أن الخوارزميات خفضت تصنيف السير الذاتية التي تحتوي على كلمة "نسائي"، كما في "نادي الشطرنج النسائي"، حتى عندما كانت المؤهلات مطابقة للمتقدمين الذكور.

فخ "**التوافق الثقافي**": تستخدم العديد من المنصات الحديثة تقييمات قائمة على الألعاب أو مقابلات الفيديو لقياس الشخصية. تشير الوثائق التي تحلل هذه الأدوات إلى أنها غالبًا ما تعاقب المتحدثين غير الأصليين أو الأفراد ذوي التنوع العصبي الذين لا تتوافق تعابير وجوههم أو نبرات صوتهم مع النموذج "القياسي" للمشاركة المبرمج في الذكاء الاصطناعي.

الاستجابات القانونية والتنظيمية في التوظيف

منظور لجنة تكافؤ فرص العمل "EEOC": تؤكد وثائق الإرشاد الأخيرة الصادرة عن لجنة تكافؤ فرص العمل "EEOC" أن "الشفافية الخوارزمية" ليست كافية. وتجادل بأنه إذا كان لأداة آلية "تأثير متباين" على مجموعة محمية، فإن صاحب العمل يكون مسؤولاً، بغض النظر عما إذا كان التحيز متعمدًا أو "مخفيًا" في التعليمات البرمجية.

فجوة الشفافية: بينما تدعي وثائق الشركات غالبًا أن نماذجها "مُصدق عليها"، يظهر تحليلنا أن هذه التصديقات غالبًا ما تتم داخليًا دون إشراف طرف ثالث، مما يؤدي إلى تضارب في المصالح بين الكفاءة المدفوعة بالربح والعدالة الاجتماعية.

الشرطة التنبؤية و"تجريم" البيانات

تمثل الشرطة التنبؤية أحد أكثر التطبيقات إثارة للجدل للحكم الخوارزمي. من خلال تحليل تقارير الشرطة وسجلات الاعتقال والوثائق القضائية، يمكننا أن نرى كيف يتم بناء "المخاطر" رياضيًا بطرق تستهدف بشكل غير متناسب المجتمعات المهمشة.

حلقة التغذية الراجعة للإفراط في المراقبة

بيانات الاعتقال مقابل بيانات الجريمة: من النتائج الهامة في تحليل الوثائق النوعي لتقارير علم الجريمة التمييز بين "الجريمة" و"الاعتقال". تعتمد الخوارزميات مثل "PredPol" أو "Compas" على بيانات الاعتقال. تشير وثائق منظمات الحقوق المدنية إلى أنه إذا تم تسيير دوريات مكثفة في حي ما، فسينتج عن ذلك المزيد من الاعتقالات، والتي تفسرها الخوارزمية بعد ذلك على أنها "نقطة ساخنة"، مما يؤدي إلى المزيد من الدوريات.

وكيل الجغرافيا: يظهر تحليلنا للمواصفات الفنية للشرطة "القائمة على المكان" أن الرموز البريدية غالبًا ما تعمل كوكيل للعرق والوضع الاجتماعي والاقتصادي.

من خلال استهداف "المناطق عالية الخطورة"، تقوم الخوارزمية بأتمتة ممارسة التمييز العنصري، وتبرر زيادة المراقبة تحت ستار الضرورة الرياضية.

تقييم المخاطر في قاعة المحكمة

دراسة حالة COMPAS: يسلط تحليل تحقيق ProPublica والاستجابات القضائية اللاحقة الضوء على عيب أساسي في خوارزميات العودية. تُظهر الوثائق أن هذه النماذج غالبًا ما تحتوي على معدلات "إيجابية كاذبة" أعلى للمدعى عليهم السود "تصنيفهم خطأً على أنهم ذوو مخاطر عالية" ومعدلات "سلبية كاذبة" أعلى للمدعى عليهم البيض "تصنيفهم خطأً على أنهم ذوو مخاطر منخفضة".

"**الصندوق الأسود**" **القضائي**: تجادل المذكرات القانونية من محامي الدفاع بأن الطبيعة الاحتكارية لهذه الخوارزميات تنتهك الحق في الإجراءات القانونية الواجبة. إذا استخدم القاضي "درجة مخاطر" لتحديد الحكم، ولكن الدفاع لا يمكنه رؤية كيفية حساب تلك الدرجة، فإن "الحكم" يصبح مرسومًا رياضيًا لا يمكن الطعن فيه بدلاً من عملية قانونية شفافة.

تجريم الإنسان

نقطة بيانات مقابل فرد: في الوثائق التي تم تحليلها، غالبًا ما يتم اختزال الأفراد إلى سلسلة من "عوامل الخطر"، مثل "العمر عند الاعتقال الأول"، "التاريخ العائلي للسجن".

محو السياق: يكشف تحليل الوثائق النوعي أن هذه الوثائق نادرًا ما تأخذ في الاعتبار العوامل المنهجية مثل الفقر، ونقص الموارد التعليمية، أو العنصرية المنهجية. تتعامل الخوارزمية مع الفرد كنقطة بيانات معزولة، متجاهلة البيئة الاجتماعية والسياسية التي تشكل مسار حياته.

الخيط المشترك للظلم الخوارزمي

تكشف مقارنة قطاعي التوظيف والشرطة من خلال تحليل الوثائق النوعي (QDA) عن منطق مشترك يتمثل في "الكفاءة على حساب المساواة".

وهم الموضوعية: في كلا القطاعين، تصور الوثائق الخوارزمية كأداة "محايدة" تعالج الحقائق ببساطة. ومع ذلك، يظهر تحليلنا أن "الحقائق"، نجاحات التوظيف السابقة أو سجلات الاعتقال السابقة، هي نفسها نتاج أنظمة اجتماعية متحيزة.

عبء الإثبات: في كل من التوظيف والشرطة، يتم نقل عبء إثبات التحيز إلى الضحية. يجب على المتقدم للوظيفة أو المدعى عليه إثبات أن "الصندوق الأسود" كان متحيزًا، وهو أمر شبه مستحيل دون الوصول إلى الكود المصدري وبيانات التدريب الخاصة.

تحويل المسؤولية: من خلال تفويض هذه القرارات إلى الذكاء الاصطناعي، تخلق المؤسسات "الشركات وإدارات الشرطة" حاجزًا للمساءلة. عندما تحدث نتيجة متحيزة، غالبًا ما يُلقى اللوم على "البيانات" أو "النموذج" بدلاً من الخيارات المؤسسية التي أدت إلى نشره.

فشـــل الخوارزميـــات في الممارسة العملية

تمثل الحالات التاليـــة "المعيار الذهبي" للتحيـــز الخوارزمي الموثـــق. من خلال تحليل الوثائق النوعـــي لتقارير ما بعد الوفاة والصحافة الاستقصائية المحيطة بهذه الأحـــداث، يمكننا رؤيـــة الآليات المحـــددة التـــي تتجلى بها تشـــوهات البيانات كضرر اجتماعي.

محرك التوظيف في Amazon

في عام 2018، تم الكشف عن أن Amazon كانت تطور أداة توظيـــف تجريبية تعمل بالذكاء الاصطناعي تـــم التخلي عنها في النهاية بسبب التحيز المتأصل على أساس النوع الاجتماعي.

الآلية: تم تدريب النظام على السير الذاتية المقدمة للشركة على مدار ١٠ سنوات. نظرًا لأن صناعة التكنولوجيا كانـــت، وما زالت، يهيمـــن عليها الذكـــور، فقـــد "تعلمت" الخوارزمية أن كونك ذكرًا كان شرطًا أساسيًا للنجاح.

مظهر التحيز: بدأت الخوارزمية في معاقبة السير الذاتية التي تضمنت كلمة "نسائي"، مثل "قائد نادي الشـــطرنج النســـائي". كما خفضـــت تصنيـــف خريجـــات كليتيـــن نسائيتين بالكامل.

يشير تحليل إغلاق المشـــروع إلى أنه حتى بعد قيام المهندســـين بتعديـــل البرامج لتكـــون "محايـــدة" لهـــذه المصطلحات المحددة، وجدت الخوارزمية "وكلاء" آخرين للنوع الاجتماعي في البيانات. يســـلط هذا الضوء على فـــخ الشـــكلية.. الاعتقاد بأنه يمكنك ببســـاطة "إصلاح" التحيز من نظام مبني على أساس متحيز.

خوارزمية Optum للرعاية الصحية

حللت دراسة نُشـــرت في مجلة Science عام 2019 خوارزمية تستخدمها Optum، شـــركة خدمات صحيـــة كبـــرى، لتحديد المرضى الذيـــن يحتاجون إلى "إدارة رعاية عالية المخاطر".

الآلية: اســـتخدمت الخوارزميـــة "تكاليف الرعايـــة الصحية" كوكيل لـــ "احتياجات الرعاية الصحية". كان المنطق هو أنه كلما زادت تكلفة الشـــخص على النظام، كلما كان أكثر مرضًا.

مظهر التحيز: بسبب الحـــواجز الاجتماعية والاقتصادية المنهجية، كان لدى المرضى الســـود تاريخيًا وصول أقل إلـــى الرعاية، وبالتالي إنفـــاق أقل على الرعاية الصحية، حتى عندما كانـــوا أكثر مرضًا من المرضى البيض. ونتيجة لذلك، صنفت الخوارزمية باســـتمرار المرضى البيـــض الأصحاء على أنهم "أكثـــر عرضة للخطر" مـــن المرضى السود الذين يعانون من أمراض مزمنة.

توضح هـــذه الحالة فخ الوكيـــل. باختيار متغير بـــدا موضوعيًا "التكلفـــة"، تجاهل المطورون الواقع الاجتماعي والسياسي بأن التكلفة ليست انعكاسًا محايدًا للصحة، بل انعكاسًا للوصول إلى الثروة والتأمين.

فضيحـــة إعانـــات رعاية الأطفال الهولندية

تضمنـــت قضيـــة إعانـــات الرعايـــة "Toeslagenaffair" فـــي هولنـــدا نظامًا خوارزميًا يســـمى "SyRI"، مؤشـــر المخاطر النظامية، مصممًا للكشف عن الاحتيال في الرعاية الاجتماعية.

الآلية: قام النظام بتحليل مجموعة واسعة من البيانات الحكومية، بما في ذلك الضرائب، والضمان الاجتماعي، والتأمين الصحي، لإنشاء "ملفات تعريف المخاطر" للأفراد الذين يُحتمل أن يرتكبوا الاحتيال.

مظهر التحيز: استهدف الخوارزمية بشكل غير متناسب الأسر ذات الدخل المنخفض والأقليات العرقية. تم اتهام آلاف العائلات زوراً بالاحتيال بناءً على "درجات المخاطر"، مما أدى إلى دمار مالي، ومصادرة المنازل، وصدمات نفسية.

أشارت الوثائق القضائية من المحكمة الهولندية التي حظرت النظام في النهاية إلى أن الخوارزمية أنشأت "طبقة رقمية دنيا". وأظهرت أنه عندما تُستخدم الخوارزميات للمراقبة الحكومية، فإنها غالبًا ما تعزز "افتراض الذنب" للمجموعات المهمشة بينما تمنح "افتراض البراءة" للأثرياء.

حد الائتمان لبطاقة Apple

في عام 2019، لاحظ رواد الأعمال والعملاء في مجال التكنولوجيا أن بطاقة Apple، التي تديرها Goldman Sachs، كانت تمنح النساء حدود ائتمان أقل بكثير من أزواجهن، حتى عندما كانت النساء يتمتعن بدرجات ائتمانية أعلى ووضع مالي أفضل.

الآلية: بينما ادعى البنك أنه لم يستخدم الجنس كمدخل، استخدمت خوارزمية "الصندوق الأسود" نقاط بيانات أخرى ترتبط بالسلوكيات المالية المرتبطة بالجنس.

مظهر التحيز: أبلغ مستخدمون رفيعو المستوى، بمن فيهم الشريك المؤسس لشركة Apple ستيف وزنياك، أن زوجاتهم تلقين حدود ائتمان أقل بما يصل إلى 20 مرة من حدودهم الخاصة على الرغم من الأصول المشتركة.

تسلط هذه الحالة الضوء على مشكلة الصندوق الأسود. نظرًا لأن عملية اتخاذ القرار في الخوارزمية كانت مملوكة، لم يتمكن ممثلو خدمة العملاء في البنك من شرح سبب اتخاذ القرار، مشيرين فقط إلى "إنها مجرد الخوارزمية". وهذا يزيل إمكانية "العدالة الإجرائية"، القدرة على الاعتراض على قرار.

الدروس المستفادة من دراسات الحالة

عندما نقوم بتجميع هذه الحالات من خلال عدسة تحليل الوثائق النوعي، تظهر عدة مواضيع متكررة تحدد "الفخ":

استمرارية الوكلاء: إن إزالة فئة محمية "مثل العرق أو الجنس" غير فعالة لأن الخوارزمية ستجد دائمًا "وكيلاً"، مثل الرمز البريدي، أو عادات التسوق، أو الإنفاق على الرعاية الصحية، يعمل كبديل لتلك الفئة.

قابلية الضرر للتوسع: على عكس مدير بشري متحيز واحد أو مسؤول قروض، يمكن للخوارزمية المتحيزة معالجة آلاف الأشخاص في الدقيقة. وهذا يزيد من حجم الظلم إلى مستوى منهجي،

مما يجعل من الصعب تحديد وتصحيح الأخطاء على أساس فردي.

فجوة المساءلة: في كل دراسة حالة، أشارت المؤسسات المعنية في البداية إلى "موضوعية" الرياضيات للدفاع عن أنظمتها. ولم يتم الاعتراف بالتحيزات إلا من خلال المراجعات الخارجية والاحتجاج العام.

نقص الوعي السياقي: تفتقر الخوارزميات إلى "الخيال الاجتماعي" لفهم سبب وجود نقطة بيانات. فهي ترى الإنفاق المنخفض على الرعاية الصحية أو فجوة في السيرة الذاتية على أنها "خطر" بدلاً من أن تكون عرضًا لعدم المساواة الاجتماعية الأوسع.

إطار "العدالة الخوارزمية".. الانتقال من الإنصاف إلى المساواة

يتطلب التحول من الإنصاف إلى المساواة تغييرًا جوهريًا في كيفية تصميمنا ونشرنا ومراجعتنا للذكاء الاصطناعي. يعتمد هذا الإطار على أربع ركائز تنقل المحادثة من "إصلاح الكود" إلى "إصلاح النظام".

من التكافؤ الإحصائي إلى الوعي السياقي

حدود "الإنصاف": تُعرّف معظم الوثائق التقنية الإنصاف بأنه "التكافؤ الإحصائي"، ضمان أن توظف الخوارزمية نفس النسبة المئوية من الرجال والنساء. ومع ذلك، إذا كانت مجموعة المتقدمين منحرفة بالفعل بسبب الحواجز المنهجية، على سبيل المثال، عدد أقل من النساء يتم تشجيعهن على دخول مجالات العلوم والتكنولوجيا والهندسة والرياضيات، فإن "الإنصاف" يدير ببساطة أعراض مشكلة أعمق.

نهج المساواة: تتطلب العدالة الخوارزمية "الوعي السياقي". وهذا يعني أن النموذج يجب أن يأخذ في الاعتبار سبب ظهور البيانات بهذا الشكل. على سبيل المثال، في تسجيل الائتمان، ستدرك الخوارزمية العادلة أن نقص التاريخ الائتماني في مجتمعات معينة هو نتيجة لـ "التخطيط الأحمر" التاريخي، وليس نقص المسؤولية المالية الفردية.

التنفيذ: يجب على المطورين تضمين "المتغيرات الاجتماعية" في تقييمات الأثر الخاصة بهم، متجاوزين "الصندوق الأسود" لفهم العوامل البيئية التي تشكل بيانات التدريب الخاصة بهم.

إجراءات التعويض الاستباقية

حدود "الإنصاف": غالبًا ما تتضمن تقنيات إزالة التحيز القياسية "العمى"، إزالة العرق أو الجنس من مجموعة البيانات. وكما رأينا في دراسات الحالة، فإن الخوارزمية تجد ببساطة وكلاء لهذه السمات، مما يحافظ على الوضع الراهن.

نهج المساواة: بدلاً من أن تكون "عمياء" عن التاريخ، يجب أن تكون الخوارزميات العادلة "واعية بالتاريخ". يتضمن ذلك "ترجيح" البيانات لمواجهة التحيزات التاريخية المعروفة. إذا تم استخدام أداة توظيف في صناعة استبعدت تاريخياً الأشخاص الملونين، فيجب معايرة الخوارزمية لتقييم "المسافة المقطوعة"، التغلب على العقبات، بقدر ما تقيم "الهيبة" التقليدية، الشهادات النخبوية.

التنفيذ: يتضمن ذلك "تضخيم البيانات الاصطناعية"، حيث يتم بشكل متعمد أخذ عينات زائدة أو "تعزيز" المجموعات الممثلة تمثيلاً ناقصًا في مرحلة التدريب

الرقابة الديمقراطية ومشاركة المجتمع

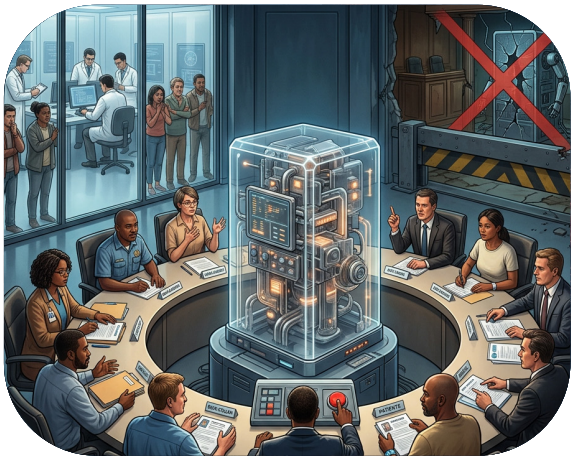
حدود "الإنصاف": حاليًا، يتم تعريف "الإنصاف" من قبل المهندسين في المختبر. الأشخاص الأكثر تأثرًا بالخوارزمية، الباحثون عن عمل، المدعى عليهم، المرضى، ليس لهم رأي في كيفية تعريف "النجاح" أو "المخاطرة".

نهج المساواة: تتطلب العدالة الخوارزمية "التصميم التشاركي". وهذا يعني أن المجتمعات المتأثرة بأداة ما يجب أن تشارك في إنشائها. إذا كانت مدينة تنشر أداة "تقييم المخاطر" للخدمات الاجتماعية، فيجب تحديد معايير "المخاطر" بالاشتراك مع الأخصائيين الاجتماعيين وقادة المجتمع، وليس فقط علماء البيانات.

التنفيذ: إنشاء "مجالس مراجعة مجتمعية" تتمتع بسلطة نقض نشر خوارزمية إذا فشلت في تلبية معايير العدالة الاجتماعية، بغض النظر عن "دقتها التنبؤية".

الرقابة الديمقراطية ومشاركة المجتمع

حدود "الإنصاف": حاليًا، يتم تعريف "الإنصاف" من قبل المهندسين في المختبر. الأشخاص الأكثر تأثرًا بالخوارزمية، الباحثون عن عمل، المدعى عليهم، المرضى،



ليس لهم رأي في كيفية تعريف "النجاح" أو "المخاطرة".

نهج المساواة: تتطلب العدالة الخوارزمية "التصميم التشاركي". وهذا يعني أن المجتمعات المتأثرة بأداة ما يجب أن تشارك في إنشائها. إذا كانت مدينة تنشر أداة "تقييم المخاطر" للخدمات الاجتماعية، فيجب تحديد معايير "المخاطر" بالاشتراك مع الأخصائيين الاجتماعيين وقادة المجتمع، وليس فقط علماء البيانات.

التنفيذ: إنشاء "مجالس مراجعة مجتمعية" تتمتع بسلطة نقض نشر خوارزمية إذا فشلت في تلبية معايير العدالة الاجتماعية، بغض النظر عن "دقتها التنبؤية".

الحق في الرفض.. الإلغاء الخوارزمي

حدود "الإنصاف": غالبًا ما يفترض الخطاب التقني أنه يمكن ويجب حل كل مشكلة باستخدام خوارزمية، طالما أننا نجعلها "عادلة".

نهج المساواة: يدرك الإطار القائم على العدالة أن بعض المهام حساسة للغاية أو أن البيانات ملوثة جدًا بحيث لا يمكن أن يكون الحكم الخوارزمي عادلاً أبدًا. في هذه الحالات، الحل الأكثر "عدلاً" هو عدم استخدام الخوارزمية على الإطلاق.

التنفيذ: تحديد "مناطق الحظر" للذكاء الاصطناعي. على سبيل المثال، يجادل العديد من دعاة العدالة بأن "الحكم التنبؤي" في المحاكم الجنائية يجب أن يُلغى تمامًا لأن البيانات التاريخية مشبعة بعمق بالعنصرية المنهجية لدرجة أنه لا يمكن لأي قدر من "إزالة التحيز" أن يجعلها عادلة حقًا.

تدقيق "المساواة"

عندما نطبق هذا الإطار على منهجية تحليل الوثائق النوعي الخاصة بنا، يمكننا تقييم الوثائق الحالية بناءً على ما إذا كانت تسعى إلى "الإنصاف" أو "المساواة".

الوثائق الموجهة نحو الإنصاف: تركز على "معدلات الخطأ" و"الإيجابيات الكاذبة" و"التحسين". إنها تتعامل مع التحيز كخلل يجب إصلاحه.

الوثائق الموجهة نحو المساواة: تركز على "ديناميكيات القوة" و"التعويض التاريخي" و"حقوق الإنسان". إنها تتعامل مع التحيز كعرض لخلل في توازن القوى يجب تفكيكه.

إن "فخ" التحيز الخوارزمي هو في النهاية فخ التوقعات المنخفضة. إذا كنا نهدف فقط إلى "الإنصاف"، فإننا نطلب من الخوارزمية أن تكون "غير متحيزة" مثل الإنسان المعيب. إذا كنا نهدف إلى المساواة، فإننا نطلب من الخوارزمية أن تساعدنا في بناء عالم أكثر عدلاً من العالم الذي تعكسه بياناتنا التاريخية.

توصيات السياسة.. أسس تشريعية للمساواة الخوارزمية

يجب أن تتجاوز السياسة الفعالة "المبادئ التوجيهية الأخلاقية"، التي غالبًا ما تكون غير ملزمة، نحو لوائح قابلة للتنفيذ تعطي الأولوية للعدالة الاجتماعية على كفاءة الشركات أو الدولة.

تقييمات الأثر الخوارزمية الإلزامية "AIAs"

على غرار تقييمات الأثر البيئي، يجب أن تكون تقييمات الأثر الخوارزمية شرطًا قانونيًا مسبقًا لأي نظام خوارزمي عالي المخاطر "التوظيف، الشرطة، الإقراض، الرعاية الصحية" قبل نشره.

عمليات التدقيق قبل النشر: يجب على المنظمات توثيق مصدر بيانات التدريب الخاصة بها، وتحديد التحيزات التاريخية المحتملة، وشرح كيفية تخفيف تلك التحيزات.

الإفصاح العام: يجب نشر ملخص لهذه التقييمات، مما يسمح لمجموعات الحقوق المدنية والباحثين بفحص منطق النظام.

الحق في "شرح ذي معنى" والانتصاف

تعتبر مشكلة "الصندوق الأسود" تهديدًا مباشرًا للإجراءات القانونية الواجبة. يجب أن تضمن السياسة أن الأفراد لديهم الحق في فهم القرارات الآلية والاعتراض عليها.

معايير قابلية الشرح: يجب أن تتطلب اللوائح أن يتلقى أي شخص حُرم من وظيفة أو قرض أو منفعة بواسطة خوارزمية شرحًا واضحًا وغير تقني للعوامل المحددة التي أدت إلى هذا القرار.

الاستئنافات بمشاركة بشرية: يجب أن يكون هناك مسار إلزامي قانونًا للمراجعة البشرية لأي قرار خوارزمي. يجب أن يكون لهذا المراجع البشري سلطة تجاوز الخوارزمية بناءً على عوامل سياقية لا يمكن للبيانات التقاطها.

الوضع القانوني للضحايا: يجب تحديث القوانين لمنح الأفراد "حق التقاضي" لمقاضاة التمييز الخوارزمي حتى لو لم يتمكنوا من إثبات "النية"، مع التركيز بدلاً من ذلك على "التأثير المتباين" للنموذج.

المراجعة المستقلة من طرف ثالث

لقد أثبت التنظيم الذاتي عدم كفايته. تُظهر تحليلاتنا النوعية لوثائق "التدقيق الداخلي" للشركات أنها غالبًا ما تعطي الأولوية لـ "الدقة" على "العدالة".

هيئات الاعتماد: يجب على الحكومات إنشاء أو اعتماد شركات مستقلة من طرف ثالث لإجراء "تدقيقات التحيز". يجب أن تتمتع هذه الشركات بإمكانية الوصول إلى الكود الخاص وبيانات التدريب الخام.

الاختبار العدائي: يجب أن يُطلب من المدققين إجراء "اختبارات عدائية"، محاولة خداع الخوارزمية عمدًا لاتخاذ قرارات متحيزة للعثور على نقاط ضعفها.

المراقبة المستمرة: نظرًا لأن الخوارزميات "تنجرف" بمرور الوقت مع استيعابها لبيانات جديدة، يجب ألا تكون عمليات التدقيق حدثًا لمرة واحدة بل متطلبًا متكررًا، على سبيل المثال، سنويًا.

أصل البيانات ومعايير الجودة

إن "الفخ التاريخي" متجذر في البيانات. يجب أن تنظم السياسة "المدخلات" بنفس صرامة "المخرجات".

ملصقات تغذية البيانات: تمامًا كما تحتوي الأطعمة على مكونات مدرجة، يجب أن تحتوي مجموعات البيانات على "ملصقات أصل" توضح بالتفصيل مصدر البيانات، ومن جمعها، وما هي الفجوات الديموغرافية الموجودة فيها.

حظر "الوكلاء السامين": يجب أن يحظر التشريع صراحةً استخدام بعض "المتغيرات الوكيلة" المعروفة بارتباطها بالفئات المحمية، على سبيل المثال، استخدام "معدلات الجريمة في الأحياء" كعامل في أقساط التأمين الفردية.

حوافز للبيانات التي تضع "العدالة أولاً": يجب على الحكومات تقديم منح وحوافز لإنشاء "مجموعات بيانات تمثيلية" تتضمن عمدًا المجموعات المهمشة لتصحيح النقص التاريخي في التمثيل.

المسؤولية والمساءلة المؤسسية

لسد "فجوة المساءلة"، يجب على القانون توضيح من هو المسؤول عندما تتسبب الخوارزمية في ضرر.

المسؤولية المطلقة للذكاء الاصطناعي عالي المخاطر: في القطاعات الحيوية مثل الرعاية الصحية والعدالة الجنائية، يجب تطبيق معيار "المسؤولية المطلقة". وهذا يعني أن المطور ومستخدم الذكاء الاصطناعي يتحملان المسؤولية عن النتائج المتحيزة، بغض النظر عما إذا كانا "ينويان" حدوث التحيز.

أحكام مكافحة الحماية: يجب منع الشركات قانونًا من استخدام ادعاءات "الأسرار التجارية"

أو "البرمجيات الاحتكارية" لحماية خوارزمياتها من التدقيق القضائي أو التنظيمي في حالات التمييز المزعوم.

حماية المبلغين عن المخالفات: يجب سن حماية أقوى للمهندسين وعلماء البيانات الذين يبلغون عن ممارسات متحيزة أو منطق تمييزي "مخفي" داخل مؤسساتهم.

دور "مدققي العدالة الاجتماعية"

أحد النتائج الرئيسية لتحليل وثائقنا النوعي هو أن الخبرة التقنية ليست كافية لتنظيم الذكاء الاصطناعي. نوصي بإنشاء دور مهني جديد: مدقق العدالة الاجتماعية.

فرق متعددة التخصصات: يجب أن تضم الهيئات التنظيمية ليس فقط علماء الكمبيوتر، بل أيضًا علماء الاجتماع والمؤرخين وعلماء القانون الذين يفهمون "الفخ التاريخي".

الرقابة التي يقودها المجتمع: يجب أن تفرض السياسة أن يكون للمجالس المحلية "لمراقبة المجتمع" مقعد على الطاولة عندما تقرر حكومة مدينة أو ولاية شراء أو نشر أداة خوارزمية.

كسر الفخ

"فخ التحيز الخوارزمي" هو خيار، وليس حتمية. من خلال التعامل مع الخوارزميات على أنها "محايدة"، نسمح لها بأن تصبح الأدوات النهائية للحفاظ على عدم المساواة المنهجية. ومع ذلك، من خلال تطبيق إطار عمل للعدالة الخوارزمية، مدعوم بتحليل وثائق نوعي دقيق ومدعوم بسياسة تركز على المساواة، يمكننا تحويل هذه الأدوات إلى أدوات للتقدم الاجتماعي.

يتطلب عصر الحكم الخوارزمي منا أن نكون أكثر يقظة، وليس أقل. يجب أن نطالب بأن تعكس تقنيتنا ليس العالم كما كان "بكل تحيزاته"، ولكن العالم كما يجب أن يكون "بكل إمكاناته للمساواة".

ما وراء ستار الحياد الرياضي

كانت الحجة الأساسية لهذا البحث هي أن "فخ التحيز الخوارزمي" ليس خللاً تقنيًا يجب "إصلاحه"، بل هو سمة أساسية للأنظمة التي تعطي الأولوية للكفاءة الرياضية على العدالة الاجتماعية. من خلال تحليلنا النوعي للوثائق للأدلة التقنية، والتدقيقات المؤسسية، والموجزات القانونية، أظهرنا أن الخوارزميات ليست مراقبين محايدين للواقع؛ بل هي مشاركون نشطون في بناء النظام الاجتماعي.

أسطورة "الصفحة البيضاء"

يكشف تحليلنا لفخاخ الشكلية، والوكيل، والتاريخية أن الخطر الأساسي للحكم الخوارزمي يكمن في "ستار الموضوعية" الخاص به. من خلال ترجمة السلوكيات البشرية المعقدة، مثل "الجدارة" في التوظيف أو "المخاطر" في الشرطة، إلى نقاط بيانات منفصلة، تجرد هذه الأنظمة السياق الاجتماعي والسياسي الذي يحدد الحياة البشرية. كما يتضح في حالات محرك التوظيف في أمازون وخوارزمية الرعاية الصحية في Optum، غالبًا ما تكون القرارات "المعتمدة على البيانات" بمثابة قناع عالي التقنية للتمييز "المعتمد على التاريخ". عندما ندرب النماذج على ماض متحيز، فإننا لا نتنبأ بالمستقبل؛ بل نؤتمت الوضع الراهن.

من "الإنصاف" إلى "المساواة"

تتمثل إحدى المساهمات الرئيسية لهذه الدراسة في التمييز بين الإنصاف الخوارزمي والمساواة الخوارزمية. فبينما يسعى الإنصاف إلى تحقيق التوازن داخل نظام قائم ومعيب، تتطلب المساواة منا التشكيك في المقاييس نفسها. ينقل إطار العدالة الخوارزمية المقترح المحادثة من "التكافؤ الإحصائي" إلى "التعويض التاريخي". ويجادل بأنه لكي تكون الخوارزمية عادلة حقًا، يجب أن تمتلك "خيالاً اجتماعيًا"، وعيًا بالحواجز المنهجية، مثل التمييز العنصري في الإسكان وعدم المساواة التعليمية، التي تشكل البيانات التي تعالجها.

الطريق إلى الأمام.. السياسة والمساءلة

يتطلب الانتقال من "الفخ" إلى "الحل" تحولاً في عبء الإثبات. تعد توصيات السياسة، بما في ذلك تقييمات الأثر الخوارزمية الإلزامية والحق في تفسير ذي معنى، ضرورية لإزالة حماية "الصندوق الأسود" التي تحمي المؤسسات حاليًا من المساءلة. يجب أن نتحرك نحو بيئة تنظيمية حيث لا يمكن استخدام "الأسرار التجارية" كدرع ضد الحق الأساسي في الإجراءات القانونية الواجبة.

تأمل أخير.. الاختيار البشري

في نهاية المطاف، فإن "فخ التحيز الخوارزمي" هو إبداع بشري، ويجب أن يكون تفكيكه مسعى بشريًا. التكنولوجيا ليست قوة مستقلة؛ إنها انعكاس للقيم والتحيزات وهياكل القوة لمنشئيها. لبناء مستقبل يعمل فيه الذكاء الاصطناعي كأداة للتحرر بدلاً من أن يكون آلية للمراقبة والإقصاء، يجب أن نطالب بأن تكون خوارزمياتنا صارمة في سعيها لتحقيق العدالة بقدر صرامتها في سعيها لتحقيق الدقة.

إن عصر الحكم الخوارزمي لا يمثل نهاية المسؤولية البشرية؛ بل يمثل بداية حدود رقمية جديدة للنضال من أجل الحقوق المدنية. يجب أن نضمن أن "حكم" الآلة يخضع دائمًا لضمير المجتمع.

المراجع

منظمة العفو الدولية. (2021). آلات كارهة للأجانب: التمييز من خلال الاستخدام غير المنظم للخوارزميات في فضيحة إعانات رعاية الأطفال الهولندية. منظمة العفو الدولية المحدودة.
https://cutt.ly/4t72IiOj

أنجوين، ج.، لارسون، ج.، ماتو، س.، وكيرشنر، ل. (2016، 23 مايو). تحيز الآلة: هناك برنامج يستخدم في جميع أنحاء البلاد للتنبؤ بالمجرمين المستقبليين. وهو متحيز ضد السود. بروبابليكا.
https://cutt.ly/Lt72IrIe

بنجامين، ر. (2019). العرق بعد التكنولوجيا: أدوات إلغاء العبودية لقانون جيم الجديد. بوليتي برس.
https://cutt.ly/ft72U7Ju

داستين، ج. (2018، 10 أكتوبر). أمازون تتخلص من أداة التوظيف السرية بالذكاء الاصطناعي التي أظهرت تحيزًا ضد النساء. رويترز.
https://cutt.ly/Zt72UVEO

المفوضية الأوروبية. (2021). اقتراح لائحة للبرلمان الأوروبي والمجلس تحدد قواعد منسقة بشأن الذكاء الاصطناعي (قانون الذكاء الاصطناعي). COM/2021/206 نهائي.
https://cutt.ly/et72UKb9

هوفمان، أ. ل. (2019). عندما يفشل الإنصاف: البيانات، الخوارزميات، وحدود خطاب مكافحة التمييز. المعلومات، الاتصال والمجتمع، 22(7)، 900-915.
https://cutt.ly/0t72USog

نايت، و. (2019، 19 نوفمبر). بطاقة Apple لم "تقصد" أن تكون متحيزة جنسيًا، ولكن كيف سنعرف ذلك على الإطلاق؟ وايرد.
https://cutt.ly/mt72UWSk

نوبل، س. يو. (2018). خوارزميات القمع: كيف تعزز محركات البحث العنصرية. مطبعة جامعة نيويورك.
https://cutt.ly/Ft72UvQp

أوبرماير، ز.، باورز، ب.، فوجيلي، س.، ومولايناثان، س. (2019). تشريح التحيز العنصري في خوارزمية تستخدم لإدارة صحة السكان. ساينس، 366(6464)، 447-453.
https://cutt.ly/At72Ufe4

أونيل، س. (2016). أسلحة الدمار الرياضي: كيف تزيد البيانات الضخمة من عدم المساواة وتهدد الديمقراطية. كراون.
https://cutt.ly/6t72Upt4

باسكوال، ف. (2015). مجتمع الصندوق الأسود: الخوارزميات السرية التي تتحكم في المال والمعلومات. مطبعة جامعة هارفارد.
https://cutt.ly/Ft72Urfh

سيلبست، أ. د.، بويد، د.، فريدلر، س. أ.، فينكاتاسوبرامانيان، س.، وفيرتيسي، ج. (2019). الإنصاف والتجريد في الأنظمة الاجتماعية التقنية. وقائع مؤتمر 2019 حول الإنصاف والمساءلة والشفافية.
https://cutt.ly/St72Y4uE

البيت الأبيض. (2025). مخطط لميثاق حقوق الذكاء الاصطناعي: جعل الأنظمة الآلية تعمل لصالح الشعب الأميركي. مكتب سياسة العلوم والتكنولوجيا.
https://cutt.ly/Dt72YHtA

الكونغرس الأميركي. (2022). S.3572 - قانون المساءلة الخوارزمية لعام 2022. الكونغرس 117.
https://cutt.ly/ht72YYWD

مركز سمت للدراسات

SMT Studies Center