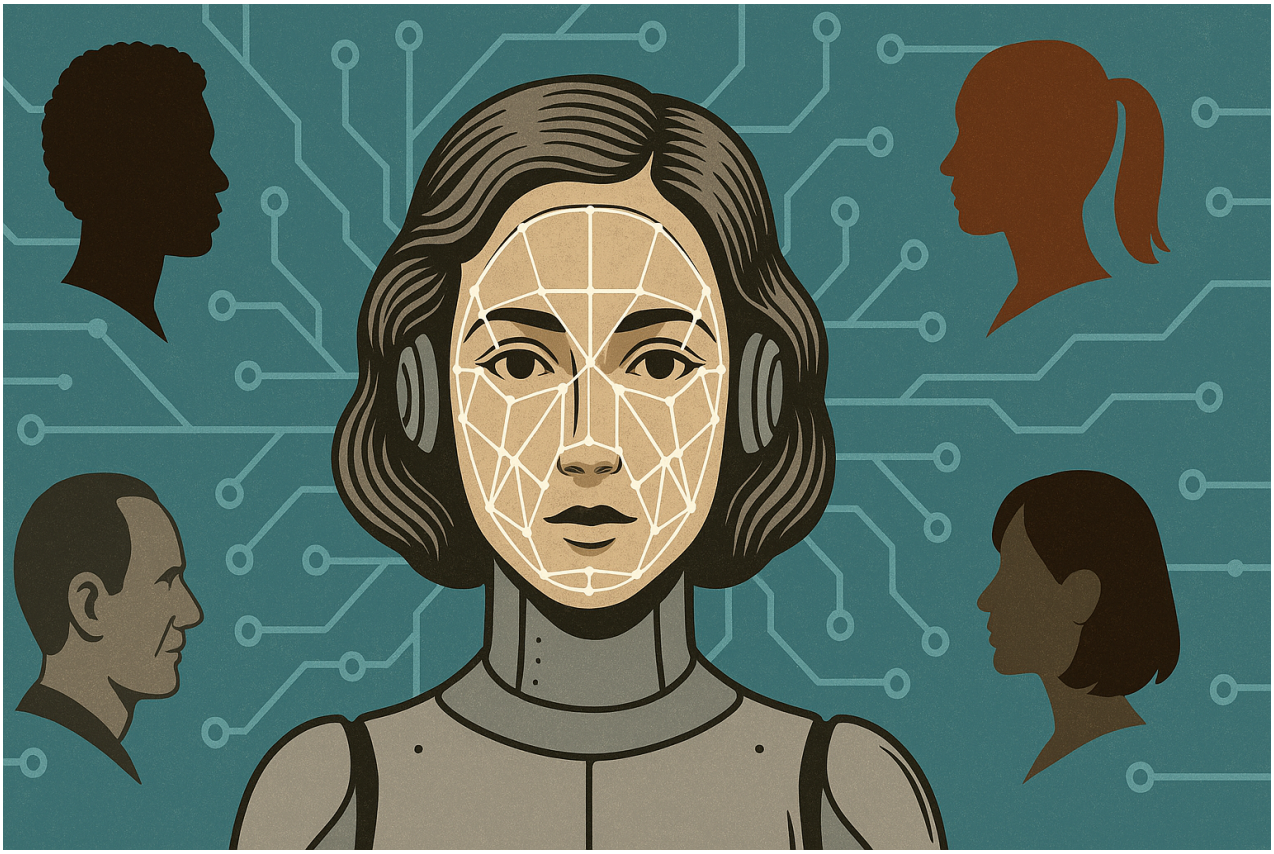
نحو عدالة خوارزمية.. كيف نواجه تحيّزات الذكاء الاصطناعي؟



تمثل تطبيقات الذكاء الاصطناعي مرآة تعكس القيم والثقافة التي صُممت في ظلها، بما تحمله من تطلعات وأحيانًا من تحيّزات غير مقصودة.

ورغم ما يُروّج له من حيادية الآلات، فإن الخوارزميات لا تنفصل عن خلفيات مطوريها ولا عن السياقات الاجتماعية التي وُلدت فيها، لهذا فإن الحدّ من انحرافات الذكاء الاصطناعي يتطلب إدراكًا عميقًا من قبل الشركات المنتجة لطبيعة المجتمعات التي تستعمل هذه الأدوات، والتأثيرات الخفية التي قد تخلّفها في سلوك الأفراد وقراراتهم.

وفي الوقت الذي يُكرس فيه الجهد لتطوير الأداء التقني وتحسين كفاءة الذكاء الاصطناعي، تظل الجوانب الأخلاقية في كثير من الأحيان مجرد هوامش، والحال أن ضمان العدالة والشفافية يتطلب منهجية متكاملة تبدأ من تصميم النماذج وصولًا إلى مراقبة أدائها بعد الإطلاق. وتشمل هذه المنهجية ضرورة إنتاج حلول متجذرة، تعكس خصوصيات المجتمعات، إلى جانب الكشف الواضح عن طبيعة البيانات التي جرى استخدامها، ومصادرها، ومواقع التحيز المحتملة فيها.

غير أن هذا التصور المثالي يصطدم بالواقع العملي الذي تحكمه السرعة والتنافس بين الشركات المطورة لهذه الأدوات والتقليل من النفقات.

ففي بيئة تكنولوجية محكومة بمنطق السوق، يغدو الالتزام بالأخلاقيات عبئًا، إلا إذا فُرض من خلال أطر تنظيمية حازمة، ولذا فإن بناء تشريعات واضحة وملزمة سيظل هو الحافز الأهم لضمان ألا يتحول الذكاء الاصطناعي من أداة للتقدم إلى آلية لإعادة إنتاج التحيزات الجندرية بصيغ أكثر تعقيدًا وخفاءً.

إعادة إنتاج التحيزات

في عام 2014، شرعت شركة أمازون في تطوير نظام توظيف قائم على الذكاء الاصطناعي، بهدف أتمتة فرز السير الذاتية وتحديد أفضل المرشحين لوظائف تقنية. استُوحي النظام من آلية تقييم المنتجات على الموقع (من نجمة واحدة إلى خمس نجوم)، وكان يُفترض به اختيار أفضل السير الذاتية من بين المتقدمين للوظائف.

لكن سرعان ما اكتشف المطوّرون أن الأداة تُظهر تحيّزًا ضد النساء، إذ تبيّن أنها تفضّل السير الذاتية التي استخدمت أنماطًا لغوية "ذكورية"، وتُقلّل من تقييم خريجات كليات النساء.

ووفق تقرير نشرته رويترز، فقد كان هذا الانحياز نتيجة تدريب الخوارزميات على بيانات من عشر سنوات، غلب عليها الذكور، مما رسّخ تصورًا بأن الرجل هو "النموذج الأفضل" للوظائف التقنية.

ورغم محاولات الشركة تعديل هذه الخوارزميات لإزالة التحيزات الظاهرة، بقيت مخاوف من ظهور أنماط تمييزية جديدة بطرق أخرى، مما دفع أمازون إلى إيقاف المشروع عام 2018.

ولم تكن تجربة أمازون هي الوحيدة في هذا السياق، حيث إن أبحاثا أكاديمية وجدت تحيزًا في تطبيق التوليدي الاصطناعي بالذكاء والرسومات الصور بإنشاء الخاص Midjourney.

وكشفت الدراسة التي أُجريت على نتائج التطبيق عن تحيّزات في الصور المولدة، وذلك بعد فحص أكثر من 100 صورة خلال ستة أشهر، حيث رُصدت عدة أنواع من التحيّزات.

فبالإضافة إلى التحيز العنصري المرتبط بكون كل الصور التي وُلّدت لمصطلحات مثل "صحفي" أو "مراسل" اقتصرت على ذوي البشرة الفاتحة، ما يعكس نقص التنوع العرقي في بيانات التدريب، والتحيز الحضري، حيث وُضعت الشخصيات دائمًا في بيئات حضرية مليئة بناطحات السحاب، حتى في غياب تحديد الموقع الجغرافي، ما يُهمّش التمثيل الريفي أو السياقات المختلفة.

بالإضافة إلى ذلك، فقد برز بشكل كبير تمييز مبني على أساس السن والجنس، حيث أظهرت الصور أشخاصًا صغار السن في الوظائف غير المتخصصة (أي الوظائف التي لا تتطلب تأهيلًا علميًا أو تقنيًا عاليًا)، بينما اقتصرت صور كبار السن على الرجال في المناصب المتخصصة فقط، مع تقديم النساء بشكل شبابي ومثالي خالٍ من التجاعيد، مقابل تسامح النموذج مع تجاعيد الرجال.

وعلى خلفية دراسة أجرتها سابقًا، حذرت اليونسكو من أن نماذج اللغات الكبيرة (LLMs)، والذكاء الاصطناعي التوليدي عمومًا، تميل إلى إنتاج صور نمطية جندرية وتوجُّهات وصفتها بـ"الرجعية"، فعلى سبيل المثال، قد ينتج نموذج AI قصة حيث يكون الطبيب دائمًا ذكرًا والممرضة دائمًا أنثى، أو يصف أدوارًا مهنية تقليدية بناءً على الجنس، مما يعزز هذه الصور النمطية بدلًا من تحديها.

يضاف إلى ذلك أن بعض مولدات صور الذكاء الاصطناعي، على غرار E-DALL 2 أظهرت تحيزًا عنصريًا وجندريًا في نتائجها، حيث قامت بربط مناصب القيادة العليا كالرؤساء التنفيذيين والمدراء بالرجال البيض وذلك بنسبة 97%.

ماذا عن المنطقة العربية؟

في السياق العربي، تبدو مسألة التحيز في الذكاء الاصطناعي وكأنها مسألة "غير موجودة"، ليس لأن النماذج الرقمية المستخدمة خالية من الانحياز، وإنما لأن الوعي بوجود هذه المشكلة لا يزال غائبًا أو مهمشًا، وذلك بالمقارنة مع السياقات الغربية، حيث جرى الكشف عن عشرات الأمثلة الموثقة لانحيازات خوارزمية ضد النساء.

ومن الملاحظ أن الساحة العربية لا تزال تفتقر إلى أدوات الرصد والتحليل القادرة على تتبع آثار هذه النماذج داخل المجتمعات.

ومن الأسباب الجوهرية لهذا الغياب هو نقص الدراسات الميدانية والتقارير الاستقصائية التي ترصد تأثير الذكاء الاصطناعي في المجالات الحيوية مثل التوظيف، والخدمات الصحية، والإعلام. فحتى الآن، لا توجد منصات بحثية أو مبادرات صحافية كافية تركز على هذا الجانب، ولا تتوفر بيانات مفتوحة تمكن من تتبع آليات اتخاذ القرارات في القطاعات الحكومية أو الخاصة.

إضافة إلى ذلك فإن غياب "البنية التحتية المعرفية والتقنية" يعمّق الفجوة، إذ إن معظم الأدوات المستخدمة في المنطقة تُطوّر خارجها، وتُطبّق محليًا دون تكييف أو فحص للسياقات الاجتماعية والثقافية التي ستعمل ضمنها. مما يعني أن احتمالات التحيز لا تُراقب أساسا، وأن الأنظمة تُعامل كصناديق سوداء لا يمكن فتحها أو مراجعة محتواها.

ومن زاوية أخرى، فإن السياق العربي يفتقر إلى ما يُعرف بـ"البيانات عن البيانات" (data-Meta)، أي القدرة على مساءلة نوعية وأصول البيانات التي تُدرّب عليها الخوارزميات. هل تمثّل النساء؟ هل تشمل مناطق جغرافية متنوّعة؟ هل تُظهر توازنًا في الأعمار، واللغات، والخلفيات الثقافية؟

الإجابة في الغالب: لا نعلم، ربما لأننا لا نسأل أصلًا

لذا، فإن السؤال الحقيقي ليس هو هل هناك تحيزات رقمية في المنطقة العربية؟ وإنما لماذا لم نبدأ بعد برصدها وتحليلها وبالتالي مواجهتها؟

وبالتالي فغياب المؤشرات لا يجب أن يُفهم كعلامة على "البراءة التكنولوجية"، لأنه يدل على ضعف أدوات النقد والشفافية التي يفترض أن تسبق أي اعتماد على التقنيات الذكية في رسم السياسات واتخاذ مختلف القرارات.

إذ لا يمكن الحديث عن ذكاء اصطناعي عادل ما لم تتوفر بيئة نقدية واعية، قادرة على فك شفرات الخوارزميات ورصد تحيزاتها، والمطالبة بتضمين العدالة الجندرية كشرط أساسي في تصميم وتطبيق النماذج الرقمية.

سبل المواجهة

مواجهة التحيز في أنظمة الذكاء الاصطناعي تبدأ بالاعتراف بوجوده وفهم آثاره على المجتمعات، مع بناء إطار أخلاقي واضح يوجّه عملية التطوير نحو الشفافية والعدالة، كما تدعو إليه مبادرات مثل Algorithm for Equality Manifesto التي تؤكد أهمية التنوع الإنساني في النماذج.

ويُعد تنويع البيانات التدريبية وتحسين جودتها خطوة محورية، إذ ينبغي أن تعكس البيانات التنوع الحقيقي في الجنس والعرق والعمر والخلفيات المختلفة، مع الالتزام بمبادئ مثلFAIR لضمان سهولة الوصول إلى البيانات وإعادة استخدامها بما يحد من التحيزات.

كما يمثل التدقيق الدوري للنماذج آلية أساسية للكشف عن التحيزات وإصلاحها، مثلما يفرض قانون مدينة نيويورك لفحص أدوات التوظيف الآلية قبل استخدامها، ويمكن للمؤسسات الاستفادة من أدوات مفتوحة المصدر مثل 36 Fairness AI لمعالجة التحيّزات في النماذج.

ويكتسب إشراك الإنسان في العملية أهمية خاصة، إذ تتيح آليات مثل BIAS-D للمستخدمين تحديد العلاقات المسببة للتحيز داخل البيانات وتعديلها، مما ينتج مجموعات بيانات أكثر حيادية.

وتعزز الحوكمة المؤسسية مواجهة التحيز عبر إنشاء لجان أخلاقيات أو مشرفين مختصين، ودمج فرق عمل متعددة التخصصات تضم خبراء قانون واجتماع ومصادر بيانات متنوعة لتوفير منظور شامل.

وإلى جانب ذلك، تلعب التشريعات دورًا مهمًا في هذا المجال، مثل المبادرات التي تقودها ولاية كاليفورنيا لوضع سياسات تحد من الممارسات التمييزية في الذكاء الاصطناعي في مجالات التوظيف

والرعاية الصحية.

وتوجد أيضًا استراتيجيات تقنية متقدمة للتقليل من التحيز، مثل Editing Concept Affine التي أظهرت فعالية في خفض الفجوة التمييزية في أنظمة التوظيف، إضافة إلى تحسين البيانات لمواجهة التحيزات في النماذج التوليدية.

إن الجمع بين الإطار الأخلاقي الواضح وتنوع البيانات والتدقيق المستمر، وإشراك الإنسان والحوكمة المؤسسية والدعم التشريعي وكذا التقنيات الحديثة، يوفر أساسًا متينًا لبناء أنظمة ذكاء اصطناعي أكثر عدالة، تعكس التنوع الإنساني وتخدمه بإنصاف.

أما عربيًا، فإن ضعف إنتاج بيانات مفتوحة ومتنوعة يُبرز الحاجة إلى مشاريع عربية تستثمر في بناء قواعد بيانات عادلة ومتنوعة، وإلى أطر حوكمة محلية تراقب عمل الذكاء الاصطناعي بما يتناسب مع الخصوصية الثقافية والاجتماعية للمنطقة.

رابط المقال: https://www.noonpost.com/339066/