

العنصرية الرقمية في أنظمة الذكاء الاصطناعي

Digital Racism in Artificial Intelligence Systems

جندولي فاطمة زهراء

مخبر قانون المؤسسة

جامعة جيلالي ليابس- سيدي بلعباس، الجزائر

Fatimadjendouli22@gmail.com

Djillali LIABES University, Sidi Bel Abbes, Algeria

بن شعيب بوسيف*

مخبر قانون المؤسسة

جامعة جيلالي ليابس- سيدي بلعباس، الجزائر

nassim.abs.bouc@gmail.com

Djillali LIABES University, Sidi Bel Abbes, Algeria

- تاريخ الإرسال: 2024/02/21 - تاريخ القبول: 2024/07/14 - تاريخ النشر: 2024/07/28

الملخص: تظهر الدراسات التي أجريت حول ظاهرة التحيز الرقمي أن العالم يشهد تحولاً هائلاً نحو تبني التكنولوجيا والذكاء الاصطناعي في جميع جوانب الحياة اليومية. يعتبر الذكاء الاصطناعي إنجازاً كبيراً في تحسين الخدمات وتسهيل الحياة، ولكنه في الوقت نفسه يطرح تحديات هامة تتعلق بحقوق الإنسان والحريات الأساسية. تشير الدراسة إلى أن الذكاء الاصطناعي يثير قضايا حساسة متعلقة بالحق في حرية التعبير والرأي والحق في الحصول على المعلومات والخصوصية والأمان والصحة والعمل. وفي هذا السياق، يتم التركيز بشكل خاص على التحيز الرقمي أو التحيز الخوارزمي، والذي ينبعث من نتائج غير حيادية تميل إلى التمييز على أساس العنصرية الرقمية. تسعى الدراسة إلى استكشاف أسباب هذه الظاهرة السريعة في عصر التكنولوجيا، وتسليط الضوء على الآليات والتدابير التي اتخذتها المؤسسات والمنظمات لمواجهة هذه الظاهرة وتقليل تأثيرها.

الكلمات المفتاحية: خوارزميات، أنظمة الذكاء الاصطناعي، التمييز، التحيز، البيانات الضخمة.

Abstract: Studies conducted on phenomenon of digital bias show that the world is witnessing a massive shift towards adopting technology and artificial intelligence in all aspects of daily life. Artificial intelligence is a major achievement in improving services and facilitating life, but at the same time it poses important challenge related to human rights and fundamental freedoms. The study indicates that artificial intelligence raises sensitive issues related to the right to freedom of expression and opinion the right to information, privacy, security, health and work. in this context, special emphasis is placed on digital bias or algorithmic bias, which stems from non-neutral results that tend to discriminate on the basis of digital racism. The study seeks to explore the causes of this rapid phenomenon in the age of technology, and to confront this phenomenon and reduce its impact.

Keywords: Algorithms- artificial intelligence systems- discriminations- bias- big data

* المؤلف المرسل: بن شعيب بوسيف

مقدمة:

في ظل التحديات الراهنة التي تواجه العالم في مجال التكنولوجيا وأتمتعدد هائل من القطاعات المختلفة، ولاسيما بعد الثورة الصناعية الرابعة التي أحدثت تحولاً جذرياً في مختلف المجالات، أصبح الذكاء الصناعي مؤشراً أساسياً لتقدم الدول وتطورها. ومع اعتماد العديد من الدول على أنظمة الذكاء الصناعي بشكل شبه كلي أصبحت تطرح قضايا قانونية وأخلاقية، خاصة فيما يتعلق بمفهوم الحيادية وعدم التحيز.

تقع المسؤولية الأخلاقية والقانونية على عاتق المبرمجين، الذين يجب عليهم برمجة الأجهزة بمراعاة القيم الأخلاقية الأساسية، مثل الحيادية وعدم التحيز، لتفادي المخاطر التي قد تتجم عن غياب هذه القيم عن عمليات اتخاذ القرار. ورغم أن هذه الأنظمة تدعم بشكل عام مبدأ الحياد، إلا أن تظهر بوضوح ظاهرة التحيز الرقمي أو العنصرية الرقمية في بعض القطاعات، خاصة تلك التي تعتمد على أنظمة الاختيار التلقائي.

تتمثل الإشكالية الأساسية لهذه الدراسة في كيفية مساهمة أنظمة الذكاء الاصطناعي في تعزيز العنصرية بين الأفراد، وكيف يمكن تحسين هذه الأنظمة لتعزيز المساواة وضمان احترام الحقوق الأساسية للفرد؟ وهي مسألة قانونية تتطلب اهتماماً فورياً وحلولاً تشريعية ذات جدية ودقة.

وتتدرج ضمن هذه الإشكالية الرئيسية تساؤلات فرعية تتمثل في:

- ما هو التحيز الرقمي وكيف يمكن تعريفه في سياق أنظمة الذكاء الاصطناعي؟
- ما هي العوامل التي تسهم في ظهور التحيز الرقمي في أنظمة الذكاء الاصطناعي؟
- كيف يؤثر التحيز الرقمي على الفرد والمجتمع بشكل عام؟
- كيف يمكن أن يؤثر التحيز الرقمي على حقوق الإنسان وحياته؟
- ما هي التدابير التي يمكن اتخاذها للوقاية من التحيز الرقمي في تطوير واستخدام أنظمة الذكاء الاصطناعي؟
- كيف يمكن تحقيق توازن بين التقدم التكنولوجي وحقوق الفرد والأخلاقيات؟

الفرضيات:

- التحيز الرقمي كتمييز غير عادل أو تمييز في نتائج أنظمة الذكاء الاصطناعي ناتج عن البيانات أو الخوارزميات.
- الأسباب الرئيسية للتحيز الرقمي هي تحيز البيانات المستخدمة في تدريب النظام وتصميم الخوارزميات.
- يكون للتحيز الرقمي آثار سلبية على المجتمع، مثل تعزيز التفاوت وتعرّض حقوق الأفراد للخطر.
- يؤثر التحيز الرقمي على حقوق الإنسان وحياته، خاصة فيما يتعلق بحقوق التعبير والخصوصية.
- هناك تدابير يمكن اتخاذها للوقاية من التحيز الرقمي، مثل تحسين جودة البيانات وتصميم خوارزميات أكثر عدالة.
- يفترض أن هناك تحديات قانونية تطرأ في سياق التحيز الرقمي تمكن من خلق التوازن بين التقدم التكنولوجي وحقوق الأفراد في المجتمع.

أهداف الدراسة

- ان دراستنا لموضوع التحيز الرقمي أو العنصرية الرقمية تتطلب تحديد أهداف واضحة لفهم الظاهرة وتأثيراتها على مستوى متعدد الأبعاد. يمكن ذكر هذه الأهداف كما يلي:
- فهم طبيعة وأسباب التحيز الرقمي في أنظمة الذكاء الاصطناعي وكيف يتم تجسيده في نتائجها
- فهم كيف يؤثر التحيز الرقمي على مختلف فئات المجتمع وكيف يمكن أن يسهم في التفاوت والظلم.
- التحقق من وجود حلول فعّالة للتحيز الرقمي، سواء كانت تقنيات أو سياسات أو تشريعات.
- وضع معايير أخلاقية توجه استخدام تقنيات الذكاء الاصطناعي بما يحقق العدالة ويحمي حقوق الأفراد.

-لمساهمة في تطوير إطار قانوني دولي ينظم استخدام وتطوير التقنيات ذات الذكاء الاصطناعي لضمان حقوق الإنسان وتعزيز العدالة الاجتماعية.

منهج الدراسة

تمت دراستنا باستخدام منهجين أساسيين هما المنهج الوصفي والمنهج الاستقرائي. حيث تم توظيف المنهج الوصفي لتحليل ووصف مشكلة التحيز الرقمي والكشف عن جذورها. في المقابل، تم اللجوء إلى المنهج الاستقرائي لفحص واستعراض النماذج المتاحة للتحيز والآليات التي تم اتخاذها للتصدي لهذه الظاهرة في السنوات الأخيرة.

خطة الدراسة

سنقوم بتقسيم الدراسة الى محورين، نتناول في (المحور الأول) إطار مفاهيمي للتحيز الرقمي وعلاقته بالذكاء الاصطناعي، وفي (مبحثان) فنتناول فيه أسباب العنصرية الرقمية في أنظمة الذكاء الاصطناعي وتداعياتها، معتمدين في ذلك على منهجية التقسيم الثنائي المعمول بها في الدراسات والأبحاث الأكاديمية.

المحور الأول: إطار مفاهيمي للتحيز الرقمي في الذكاء الاصطناعي

سنتناول في هذا المحور مفهوم الذكاء الاصطناعي (أولاً) ثم تعريف التحيز الرقمي وأنواعه (ثانياً).

أولاً - مفهوم الذكاء الاصطناعي وأهميته

تقتضي منا الدراسة التعرف على مفهوم الذكاء الاصطناعي كعلم حديث ثم التعرض بعد ذلك الى أهميته في حياة البشر كوافد جديد جراء الثورة الصناعية الرابعة.

1- تعريف الذكاء الاصطناعي:

انطلاقاً من تنامي دور التقنيات التكنولوجية الحديثة في شتى مجالات الحياة بصورة ملحوظة، نجد أن أكثر ما يخدم هذا الامر ويؤيده هو ظهور أنظمة الذكاء الاصطناعي وتنامي دورها في الكثير من المجالات الحياتية. وأول ما يثير الذهن بشأن هذه الأنظمة الحديثة، هو بيان تعريفها. وتتأتى الأهمية

العملية لأنظمة الذكاء الاصطناعي من منطلق اعتبارها من ضرورات الحياة العصرية، لكونها تسهم في تسهيل القيام بالعديد من الأمور بأداء يفوق الأداء البشري¹.

وضع العديد من الفقهاء تعريفات مختلفة للذكاء الاصطناعي، فعرّفه البعض بأنه: "علم وتقنية قائمة على عدد من المجالات المعرفية، مثل علوم الحاسبات الآلية والرياضيات، والأحياء، والفلسفة، والهندسة، والتي تستهدف تطوير وظائف الحاسبات الآلية لتحاكي الذكاء البشري"². كما عرف بأنه: "جزء من علوم الحاسب الآلي يهدف إلى تصميم أنظمة ذكية، تعمل على جعل الحاسب الآلي يمثل ويحاكي التفكير الإنساني، وبعض قدرات السلوك الإنساني، ويعطيها ذات الخصائص التي نعرفها بالذكاء في السلوك البشري". ويرى بعض العلماء أنه: "التيار العلمي والتقني الذي يضم الطرق والنظريات والتقنيات التي تهدف إلى إنشاء آلات قادرة على محاكاة الذكاء البشري"³.

كما تم تعريف الذكاء الاصطناعي بأنه أنظمة برمجيات أو أجهزة من تصميم بشري ذات أهداف معقدة، وتنشط في العالم الحقيقي أو الرقمي من خلال إدراك البيئة بواسطة تلقي مجموعة من المعلومات، ومن خلال تفسير البيانات المهيكلة أو غير المهيكلة المجمعة، وتطبيق تحليل على المعارف أو معالجة المعلومات المستمدة من تلك البيانات، وتقرير الإجراء أو الإجراءات الأفضل الواجب اتخاذها من أجل تحقيق هدف معين. ويمكن لنظم الذكاء الاصطناعي إما استخدام قواعد رمزية، أو تعلم نموذج رقمي، كما يمكنها أيضًا تكيف سلوكها من خلال تحليل كيفية تأثر البيئة بإجراءاتها السابقة⁴.

ومما سبق يمكن القول بأن الذكاء الاصطناعي هو عبارة عن العلم الذي يهدف إلى تصميم أنظمة ذكية، من شأنها أن تجعل الحاسب الآلي يحاكي التفكير البشري ويتعامل بذات القدرات البشرية، وذلك من خلال تغذيته بالبيانات والمعلومات الضخمة، أو من خلال التعلم الذاتي.

2- أهمية الذكاء الاصطناعي

أصبح الذكاء الاصطناعي موضوعًا رئيسيًا يظهر في محادثاتنا الشخصية والمهنية، وفي المناقشات السياسية، وفي المؤتمرات الصناعية والرقمية، وما إلى ذلك. كما يعتمد الذكاء الاصطناعي

¹ - محمد فتحي محمد إبراهيم، التنظيم التشريعي لتطبيقات الذكاء الاصطناعي، مجلة البحوث القانونية والاقتصادية، العدد 81 (سبتمبر 2022)، كلية الحقوق، جامعة المنصورة، جمهورية مصر العربية، ص 1021.

² - عبد اللاه إبراهيم الفقي، الذكاء الاصطناعي والنظم الخبيرة، دار الثقافة للنشر والتوزيع، 2012، ص 17، د/أشرف السعيد احمد، الذكاء الاصطناعي والنظم الخبيرة، أكاديمية الشرطة، معهد بحوث الشرطة، ص 4.

³ محمد فتحي محمد إبراهيم، مرجع سبق ذكره، ص 1022

⁴ - Commission européenne, Lignes directrices en matière d'éthique pour une intelligence artificielle digne de confiance, 8 avril 2019, p9 et 143.

على إنشاء وتطبيق الخوارزميات التي يتم تنفيذها في بيئة حوسبة ديناميكية. هدفها هو السماح لأجهزة الكمبيوتر بالتفكير والتصرف مثل البشر. عندما نتحدث عن هواتفنا الذكية، وأجهزة الكمبيوتر لدينا، والأشياء المتصلة بنا، فإننا نشير بالفعل إلى الذكاء الاصطناعي، مما يعني أنه جزء من حياتنا اليومية. إنه منتشر في كل مكان، مما يعني أنه حاضر ويتدخل في جميع المجالات تقريباً⁵. ومن بين الأمثلة حول استخداماته اليومية:

- التسوق والإعلان عبر الإنترنت حيث يتم استخدام الذكاء الاصطناعي بشكل متكرر عند تقديم توصيات مخصصة للمستخدمين.
- البحث على الانترنت اذ تتطور محركات البحث لأنها تجمع تدفقاً هائلاً من البيانات المقدمة من المستخدمين، من أجل تزويدهم بنتائج أكثر صلة.
- -المساعدات الرقمية الشخصية تستخدم الهواتف الذكية الذكاء الاصطناعي لتقديم منتج ملائم ومخصص قدر الإمكان. يجب المساعدون الافتراضيون (الموجودون الآن في كل مكان) على الأسئلة ويقدمون التوصيات ويساعدون في إدارة المهام اليومية.
- الترجمة الآلية تعتمد برامج الترجمة، المعتمدة على اللغة المكتوبة أو المنطوقة، على الذكاء الاصطناعي لتقديم ترجمات محسنة باستمرار. ينطبق هذا الأسلوب أيضاً عند إنشاء الترجمات تلقائياً.
- ربط المنازل والمدن والبنية التحتية تقوم بعض أجهزة تنظيم الحرارة بتحليل سلوكنا لتوفير الطاقة بشكل أفضل، بينما يأمل المخططون الحضريون للمدن المتصلة في تقليل الاختناقات المرورية وإدارة حركة المرور بشكل أفضل
- على الرغم من أن السيارات ذاتية القيادة لم تصبح هي القاعدة بعد، فإن سياراتنا تستخدم بالفعل ميزات السلامة القائمة على الذكاء الاصطناعي. على سبيل المثال، دعم الاتحاد الأوروبي تمويل أجهزة الاستشعار الأوتوماتيكية VI-DAS التي تكتشف المواقع والحوادث التي يحتمل أن تكون خطيرة .
- الأمن الإلكتروني يمكن للأنظمة التي تدعم الذكاء الاصطناعي أن تساعد في اكتشاف الهجمات الإلكترونية والتهديدات الأخرى ومكافحتها، وذلك باستخدام التدفق المستمر للبيانات للكشف عن الأنماط وتتبع مصدر الهجمات⁶.

⁵-Inspace Institute, Quel est l'importance DE L'intelligence artificielle dans le monde ,publié le 08-06-2021,<https://www.inspace-institute.com/quelle-est-limportance-de-lintelligence-artificielle-dans-le-monde/>, consulté le 10-02-2024 à 19h25mn.

⁶- Thème parlement européen, Intelligence artificielle : définition et utilisation,L'intelligence artificielle (IA) est en passe de devenir la « technologie clé de l'avenir ». Mais qu'entend-on exactement par « IA » et comment affecte-t-elle notre quotidien?,publié le 07-09-2020,dernière mise à jour 07-06-2023 à 16 :17.,<https://www.europarl.europa.eu/topics/fr/article/20200827STO85804/intelligence-artificielle-definition-et-utilisation>, consulté le 14/02/2024 à 17h12mn.

○ الذكاء الاصطناعي ضد كوفيد-19 يتيح الذكاء الاصطناعي مكافحة فيروس كورونا، من خلال استخدامه في المطارات أثناء إنتاج الصور الحرارية وكذلك في حالات أخرى. وفي المجال الطبي، يمكن للذكاء الاصطناعي اكتشاف العدوى عندما تخضع رئة المرضى للتصوير المقطعي بالكمبيوتر. كما أتاح أيضاً جمع البيانات من أجل مراقبة تطور حالات العدوى.

ثانياً - مفهوم التحيز الرقمي وأنواعه

سنتناول تعريف التحيز الرقمي (1)، ثم أنواعه (2)

1- تعريف التحيز الرقمي:

التحيز في الذكاء الاصطناعي يمكن تعريفه كظاهرة تميز غير مقصود تظهر في نتائج نظام الذكاء الاصطناعي، والتي قد تؤدي إلى اتخاذ قرارات غير عادلة أو غير متوازنة⁷.

كما يمكن تعريفه كتشويه في نتائج خوارزميات التعلم الآلي ناتج عن افتراضات متحيزة خلال مراحل تطوير الخوارزمية، مما يعكس تمييزاً اجتماعياً وتحيزاً ضد فئة محددة. قد ينشأ هذا التحيز نتيجة لتفضيل أو إهمال معين في بيانات التدريب، أو بسبب افتراضات متحيزة في تصميم الخوارزمية نفسها، أو حتى بسبب تحيز المطورين أثناء عملية التطوير. يعمل المجتمع العلمي والتكنولوجي على التعامل مع هذا التحيز من خلال تبني مبادئ التصميم العادل وتحسين عمليات جمع البيانات وتدريب النماذج لضمان نتائج ذكية أكثر عدالة وتوازناً⁸.

ومع انتشار وتوسع تطبيقات الذكاء الاصطناعي في مختلف القطاعات والمجتمعات، أصبح تأثير التحيز واضحاً بشكل واسع النطاق. ومع تزايد هذا التأثير، تنامت الحاجة إلى نظم عادلة قادرة على اتخاذ قرارات أو دعم عمليات اتخاذ القرار تتمتع بالإنصاف وتخلو من أشكال العنصرية والتحيز. هذا التطور يعكس الحاجة المتزايدة إلى إقامة نظم تكنولوجية تعكس قيم المساواة والعدالة في مجتمعاتنا المتنوعة.

2- أنواع التحيز الرقمي:

يعدُّ التحيز الخوارزمي ظاهرةً جديدةً في عالم التقنية، حيث أثارت مخاوف واسعة من إمكانية تعرُّض الأفراد حول العالم لأضرارٍ في حياتهم الشخصية والعملية والاجتماعية، فضلاً عن أوضاعهم

⁷-Amr Okasha, Bias in Artificial intelligence, www.linkedin.com, publié le 16 juillet 2023, consulté le 01/01/2024 à 22h30mn.

⁸- مسار، كيف يمكن التغلب على تحيز الذكاء الاصطناعي... تقنيات و أدوات ، تاريخ النشر يوم 7 مارس 2022 ، <http://massar.net> ، تاريخ الزيارة 2024/01/01 ، 22ما56د.

القانونية نتيجة لانتشار أنظمة الذكاء الاصطناعي في مجالات متعددة. يتضمن ذلك التوظيف، ومكافحة الجريمة، والائتمان، والمعاملات المالية، والبنكية، والإفراج المشروط عن المسجونين، وغيرها. كما أظهرت الأبحاث أن النساء والأقليات، وبعض العقائد والتنوعات الاجتماعية، يكونون عرضةً بشكل رئيسي للتحيز الخوارزمي. وبناءً على هذا الأساس، يُعدُّ التحيز الخوارزمي عاملاً مُساهمًا بشكل متزايد في ظاهرة العنصرية والتحيز الذي قد تظهر به خوارزميات الذكاء الاصطناعي. يمكن أن تظهر أشكال عديدة من "التحيز الخوارزمي" في نتائج الذكاء الاصطناعي والأنظمة المستقلة يمكن ذكر الشائعة منها:

1-2- التحيز الجنسي (Genderbias): هو عندما يتم تفضيل جنس واحد على جنس آخر أو عدم تمثيله في النتيجة على الإطلاق، ومن الأمثلة المعروفة على ذلك خوارزمية التوظيف التي تفضل الرجال على النساء في أنواع معينة من الأدوار

2-2- التحيز العنصري (Racial bias): يحدث عندما يتم تفضيل مجموعات عرقية معينة على مجموعات عرقية أخرى. على سبيل المثال، يمكن أن يأتي هذا من الخوارزميات المستخدمة في نظام العدالة الجنائية لتحديد احتمالية تكرار ارتكاب جريمة ما، ومن الممكن أن تختار الخوارزمية مجموعة عرقية واحدة على أخرى بسبب عرقها.

3-2- التحيز العمري (Agebias): يحدث عندما يتم تفضيل مجموعات الأشخاص من فئات معينة على فئة عمرية أخرى. يمكن أن يكون هذا نتيجة ثانوية لتمثيل البيانات، حيث قد تنتج الأجيال الأكبر سناً كمية مختلفة من المعلومات اللازمة لتمثيلها مقارنة بالأجيال الأصغر سناً على قدم المساواة⁹.

4-2- التحيز التأكيدي (Confirmationbias): يحدث عندما تنتج الخوارزمية حلقة تعليقات من نتائجها السابقة.

5-2- تحيز التمثيل (Representationbias): يحدث عندما لا يتم تمثيل مجموعة من الأشخاص في بيانات الخوارزمية. ومن الأمثلة على ذلك عدم تمثيل السكان الأصليين بنفس الدرجة مثل المجموعات الأخرى من الأشخاص في مجموعات البيانات المختلفة التي ترب خوارزميات الذكاء الاصطناعي.

6-2- انجراف المفهوم (Concept drift): يحدث عندما لا يتكيف النموذج الخوارزمي مع الأحداث غير المتوقعة التي قد يواجهها شخص ما. من الأمثلة على ذلك الخوارزمية التي تحاول التنبؤ

⁹JefferyRecker, Differents types of biasfound in Artificial Intelligence Algorithms , publié le 17 janvier 2023 <https://medium.com/@jeffery-recker/different-types-of-bias-found-in-artificial-intelligence-algorithms-805106d992cf> , consulté le 03/01/2024 à 23h15mn.

بقرارات الشراء الخاصة بالفرد ولكنها تقشل في التغيير بعد انتقاله الى مدينة جديدة بشكل صحيح. يحصل هذا الشخص على متاجر معن عنها من المدينة التي كان يعيش فيها سابقا بدل من المدينة التي يعيش فيها حاليا.

2-7- تحيز الخصوصية: (Privacybias): يحدث عندما لا يتم تمثيل مجموعة من الأشخاص في بيانات النظام الآلي لأنهم اختاروا عدم جمع بياناتهم. ومن الأمثلة على ذلك عدم قيام منصة التواصل الاجتماعي بجمع البيانات من مجموعة من المشاركين وتقديم نتائج مختلفة على منصتهم نتيجة لذلك.

2-8 - تحيز الإعاقة (Disabilitybias): يحدث عندما تقشل الخوارزمية في المستخدمين ذوي الإعاقات العقلية أو الجسدية، مثل ضحايا السكتة الدماغية، والأشخاص المصابين بالتوحد. ومن الأمثلة على ذلك خوارزمية التعرف على الوجه التي تحاول قراءة المشاعر الإنسانية والفشل في التعرف على شخص يكافح من أجل التعبير عن مشاعر مختلفة بشكل صريح.

2-9 - تحيز اللغة: (Languagebias): يحدث عندما تقشل الخوارزمية في معاملة الأشخاص الذين لديهم فهم مختلف للغة الأصلية للخوارزمية مثل المتحدثين الأصليين الآخرين. مثال على ذلك هو قيام خوارزمية بتقييم الأوراق الجامعية حول موضوع ما، والفشل في الفهم الكامل لما يقوله كاتب غير أصلي ومعاقتهم بشكل غير عادل.

2-10 - التحيز الإقليمي (Regionalbias): يحدث عندما لا يتم تمثيل مجموعات معينة بدقة في بيانات الخوارزمية. مثال على ذلك الخوارزمية التي تتنبأ بما إذا كان ينبغي الموافقة على شخص ما للحصول على قرض أم لا بناء على البيانات التاريخية، والتي تقشل في فهم السياق التاريخي لسبب عيش بعض الأشخاص في المكان الذي يعيشون فيه، وفي النهاية تعاقب هؤلاء الأشخاص كنتيجة لذلك.

2-11 - انحياز الحداثة (Recencybias): يحدث عندما تفضل الخوارزمية البيانات الأحدث على البيانات التاريخية. ومن الأمثلة على ذلك الخوارزمية التي تحاول التنبؤ بالاتجاهات في الأسواق المالية وتفضل البيانات الحية الحالية على البيانات التاريخية لتلك الأسواق نفسها.

2-12 - التحيز الثقافي (Culturebias): يحدث عندما تفضل الخوارزمية ثقافة ما على ثقافة أخرى. ومن الأمثلة على ذلك فشل التعرف على الوجه في تحديد الشخص الذي يستخدم تسريحات الشعر الأكثر تمثيلا لتراثه الثقافي مقارنة بتسريحات الشعر الشائعة الاستخدام والأكثر تمثيلا لبيانات التدريب الخاصة بالخوارزمية.

إضافة الى ما ذكرناه من أنواع للتحيزات الرقمية التي تظهر في نتائج أنظمة الذكاء الاصطناعي يمكن العثور على تحيزات أخرى في العديد من الاشكال المختلفة وجميع أنواع الخوارزميات. وغيه فان أفضل طريقة لتقليل المخاطر المرتبطة بالتحيز الخوارزمي هي فهم أنواع التحيز المختلفة والمتميزة التي يمكن أن تحدث، والتفكير بشكل استباقي في الكيفية التي يمكن بها للخوارزمية أن تلحق الضرر بشخص ما، ثم وضع التدابير اللازمة لمنعها او التقليل منها¹⁰.

المحور الثاني: أسباب العنصرية الرقمية في الذكاء الاصطناعي وتدابير الحد منها

سنتعرض في هذا المحور الى جزئيتين هامتين تتمثل الأولى في معرفة أسباب العنصرية الرقمية في أنظمة الذكاء الاصطناعي (أولا) بينما تتضمن الجزئية الثانية أهم التدابير الوقائية والاحترازية التي يمكن تفعيلها للحد او التقليل من هذه الظاهرة (ثانيا).

أولا- أسباب العنصرية الرقمية في أنظمة الذكاء الاصطناعي

العنصرية الرقمية تعكس ظاهرة تتجلى في العالم الرقمي، حيث يحدث التمييز والاضطهاد بناءً على العرق أو اللون أو الدين وغيرها من الخصائص ذات الصلة في السياق الرقمي. مع انتشار استخدام الشبكات الاجتماعية وتوسع التواصل الرقمي في حياة الأفراد، أصبحت تجارب العنصرية تظهر بشكل متزايد في هذا الفضاء. كماتعبر العنصرية الرقمية عن التحديات نفسها التي تواجهها المجتمعات في العالم الواقعي، مثل رفض الآخر والتمييز العنصري، ولكنها تظهر في سياق الأنشطة الرقمية. كما يمكن أن تظهر هذه الظاهرة في خوارزميات البحث، حملات التسويق الرقمي، أو حتى في التفاعلات الاجتماعية عبر الإنترنت. في هذا السياق، يشير مصطلح العنصرية الرقمية إلى التحديات التي يمكن أن يواجهها الأفراد على الإنترنت بسبب خصائصهم الشخصية، وكيف يمكن للتكنولوجيا أن تعكس أو تعزز التحيزات الموجودة في المجتمعات.

العنصرية الرقمية في الذكاء الاصطناعي تشير إلى وجود تحيز أو تفضيل غير عادل في نظم الحوسبة، سواء كان ذلك في البيانات المستخدمة أو في الخوارزميات المستخدمة لاتخاذ القرارات. يمكن أن يظهر هذا التحيز في تأثيرات سلبية على فئات معينة من الأفراد أو المجتمعات، سواء على صعيدي الوصول إلى التكنولوجيا أو في تكامل الخدمات الرقمية.

هناك عدة أسباب قد تؤدي إلى العنصرية الرقمية في أنظمة الذكاء الاصطناعي، وتشمل هذه الأسباب:

¹⁰-JefferyRecker,ibid.

1- تحيز البيانات: من أهم التحديات في مجال الذكاء الاصطناعي التحيزات المعرفية إذ تنطلق في كثير من الأحيان من البشر أنفسهم الذين يقومون بتصميم النماذج وتدريب الخوارزميات. إذا كان لدى المطورين أو المهندسين تحيزات غير واعية أو مفترضة، فإن هذه التحيزات قد تنعكس في تصميم وأداء النموذج.

على سبيل المثال، إذا قام المطورون بتدريب نموذج على مجموعة معينة من البيانات التي تحتوي على تحيزات غير عادلة أو غير متوازنة، فإن النموذج قد يتعلم هذه التحيزات ويظهرها في قراراته. كما أن التحيزات الموجودة في البيانات التدريبية يمكن أن تؤثر على أداء النموذج على نطاق واسع¹¹.

2- تصميم الخوارزميات: لا يمكن و بنسب كبيرة تحقيق العدالة و الانصاف و الابتعاد عن التحيز اثناء تطوير أنظمة الذكاء الاصطناعيلعدة أسباب؛ ربما تكون أهمها هو ادخال نماذج التعلم الآلي بالبيانات التي تُجمع من العالم الحقيقي، وبالتالي فإن أكثر الأنظمة دقة يمكن أن تتعلم أو تكرر التحيزات الموجودة مسبقًا في هذه البيانات، تحيزات قائمة سواء على أساس العرق أو الجنس أو الدين أو أي خصائص أخرى. يمكن أيضًا أن يكشف النظام عن نقاط عمياء غير مقصودة بعد إطلاقه، هذه النقاط قد تكون إشكاليات تظهر قبل أو أثناء أو بعد تطوير نظام الذكاء الاصطناعي نتيجة وجود تحيزات أو أحكام مسبقة أو تفاوتات هيكلية في المجتمع، وتحدث حتى مع تدريبه واختباره بشكل صارم. على سبيل المثال، عندما يتم تدريب نظام على الاعتراف بالأصوات البالغة، قد يفتقر إلى التعرف الفعال على الكلمات والعبارات العامية التي يستخدمها المراهقون. هذا النوع من التحيز يشير إلى ضرورة تحسين تنوع بيانات التدريب واستخدام تقنيات التعلم العميق لتحقيق أداء أكثر عدالة وشمولاً. بالإضافة إلى ذلك، يجب أن يتم التحقق الدوري والتحديث المستمر للنظام لمواكبة التطورات اللغوية وتعبيرات الفئات المختلفة.

3- قلة التنوع في الفريق المطور: قلة التنوع في فريق تطوير خوارزميات الذكاء الاصطناعي يمكن أن يسهم في ظهور التحيز الرقمي في النظام. إذ أن الأفراد يمكن أن يأتوا من خلفيات ثقافية واجتماعية محددة، وقد يكون لديهم فهم محدود لتجارب الأفراد الآخرين. هذا قد ينعكس في البيانات التي يتم استخدامها وفي الخوارزميات التي تصمم.

4- قصور البيانات: نقص البيانات يمكن أن يؤدي إلى التحيز في نظم الذكاء الاصطناعي. إذا كانت البيانات المستخدمة في تدريب النموذج غير متوازنة أو غير كافية، فإن النموذج قد يظهر تحيزات في أدائه. في دراسات ميدانية قامت بها الحكومة الاتحادية في الولايات المتحدة الأمريكية، أن أنظمة التعرف على الوجه المعتمدة من وكالات إنفاذ القانون والمطارات و الموانئ و الجمارك و في أي جهة

¹¹- Thomas Leduc, Intelligence artificielle : savoir expliquer les biais, publié le 28 aout 2021, <https://golem.ai/fr/blog/entreprise/biais-algorithmes-ia>, consulté le 03 janvier 2024 à 20h18mn.

تعتمد هذه التقنيات تحتوي على أخطاء ذات الدلالات العنصرية. فبحسب الدراسة ، أخطأت التقنية في تعريف الأشخاص الملونين أكثر من البيض ، و كان الأشخاص الأمريكيون من أصل إفريقي و آسيوي أكثر عرضة للتمييز بشكل خاطئ 100 مرة من الرجال البيض، اعتمادا على الخوارزمية و نوع البحث ، بينما تم التعرف على وجوه النساء الأمريكيات من أصل إفريقي بشكل خاطئ في كثير من الأحيان ¹².

ثانياً تدابير احترازية للتقليل من العنصرية الرقمية أو الحد منها

سمحت أنظمة الذكاء الاصطناعي للأفراد حول العالم بتجربة تحديات ومهارات جديدة، حيث تم استخدامها بشكل واسع في مجموعة متنوعة من الميادين. بدءًا من ترشيحات الكتب والبرامج التلفزيونية وصولاً إلى تنفيذ مهام أكثر تعقيداً مثل اختيار المتقدمين للوظائف المختلفة، وحتى المهام ذات الطابع الحساس مثل التنبؤ بتفشي الأمراض.

تتطلب أنظمة الذكاء الاصطناعي كميات كبيرة من البيانات لتحسين دقتها. يتم جمع وتنظيف البيانات ثم معالجتها باستخدام خوارزميات ذكية، مما يتيح للنظام فهم أنماط وخصائص البيانات. هذه العمليات تسهم في تطوير قدرة الآلة على أداء المهام المصممة لها.

مع الانتشار الواسع لتطبيقات الذكاء الاصطناعي في مختلف المؤسسات والمجتمعات ازداد تأثير تحيزها بشكل واسع النطاق، ما أدى الى حتمية طلب المجتمعات الى أنظمة عادلة قادرة على اتخاذ قرارات - أو المساعدة في اتخاذها- تتسم بالإنصاف ومُجَرَّدة من أشكال العنصرية والتحيز، و منه تطوير أنظمة الذكاء الاصطناعي بشكل عادل و خال من التحيز يشكل تحدٍ كبير، حيث يعتمد هذا التحدي على تدفق البيانات من العالم الحقيقي. يمكن للنماذج أن تتعلم وتعكس التحيزات الموجودة في البيانات، مثل التحيزات العرقية أو الجنسية. يمكن أيضًا أن تظهر نقاط عمياء غير متعمدة في أداء النظام، مما يجعله غير فعال في بعض السيناريوهات. يتعين على المطورين الحذر في جمع البيانات وتحسين عمليات التدريب، وتطوير آليات لتصحيح التحيزات المحتملة في النماذج. الالتزام بمبادئ الأخلاق في تطوير الذكاء الاصطناعي يعد أمرًا أساسيًا لضمان تحقيق تطبيقات عادلة وشاملة. على سبيل المثال، في سياق تدريب نظام الذكاء الاصطناعي على التعرف على أصوات البالغين، يظهر العدالة والشمول في هذا النطاق. ومع ذلك، قد يظهر التحيز عندما يتعامل النظام مع الكلمات أو العبارات العامة المستخدمة بشكل شائع بين المراهقين. هذا يبرز أهمية تحسين تنوع بيانات التدريب واستخدام تقنيات التعلم العميق ¹³.

¹² - طه الراوي، عنصرية التكنولوجيا.. عن أسباب أشكال التحيز الرقمي ، تاريخ النشر 12 يونيو، 2020، <http://www.noonpost.com>

، تاريخ الزيارة 03 جانفي 2024، 22 سا 30د.

¹³ - مسار، مرجع سبق ذكره.

1- ممارسات للحد من تحيز الذكاء الاصطناعي

قصد التقليل من مشكلة تحيز الذكاء الاصطناعي نشرت شركة Innodata¹⁴ المتخصصة في هندسة البيانات بعض الحلول متمثلة في خمسة ممارسات للتقليل و الحد من العنصرية الرقمية لأنظمة الذكاء الاصطناعي و بالتالي الحصول على نماذج تعلم آلي تتوفر على قدر كبير العدل و الانصاف متمثلة فيما يلي :

1-1- اختيار مجموعة البيانات: تعتبر خطوة حاسمة في التقليل من تحيز الذكاء الاصطناعي وتحقيق التعلم الآلي الشامل والعدل. البيانات التي يتم استخدامها في تدريب نماذج الذكاء الاصطناعي يجب أن تكون شاملة ومتنوعة لضمان تمثيل جميع الفئات والسيناريوهات الممكنة:

- **التنوع الجغرافي والثقافي:** يجب أن تشمل مجموعة البيانات مختلف الثقافات والخلفيات الاجتماعية والجغرافيات. يساعد هذا في ضمان تمثيل متنوع للمجتمعات المختلفة التي يتم استهدافها بواسطة النموذج.

- **توازن الفئات:** يجب أن يكون هناك توازن في عدد البيانات الممثلة لكل فئة داخل مجموعة البيانات. على سبيل المثال، إذا كان النموذج يستهدف الجنسين، فيجب أن يكون هناك تمثيل متساوٍ للذكور والإناث في البيانات.

- **الشمولية الزمنية:** يجب أن تغطي مجموعة البيانات فترات زمنية متعددة لضمان تمثيل مختلف الظروف والسياقات التي قد تتغير مع مرور الوقت.

- **توفر البيانات الصافية:** يجب أن تكون البيانات خالية من التحيز والمعالجة بعناية لتجنب تضليل النماذج الذكية.

باختيار مجموعة بيانات شاملة ومتنوعة، يمكن للشركات والمنظمات تحقيق نماذج ذكاء اصطناعي أكثر عدالة وشمولية، مما يساهم في تعزيز الثقة وتقليل التحيز في التطبيقات الذكية.

1-2- تنوع الفريق: تنوع الفريق هو عنصر رئيسي يمكن أن يحد من التحيز في نماذج التعلم الآلي ويسهم في إنشاء مجموعات بيانات متوازنة وممثلة:

¹⁴- شركة . Innodata المعروفة سابقًا باسم . Innodatalysogen، هي شركة أمريكية تقدم خدمات العمليات التجارية والتكنولوجيا والخدمات الاستشارية. توفر الشركة أيضًا منتجات تهدف إلى مساعدة العملاء على إنشاء المعلومات الرقمية وإدارتها واستخدامها وتوزيعها. https://ae.linkedin.com/company/innodata-inc-?trk=ppro_cpro، تاريخ الزيارة 10 جانفي 2024، ص 30 د.

- فهم أعمق للسياقات الثقافية:توظيف فريق متنوع يعني وجود خبرات وثقافات متعددة. هذا يساعد في فهم أفضل للسياقات الثقافية المختلفة الممكن أن تظهر في البيانات. فريق يمتلك خلفيات متنوعة يمكنه التعرف على التحيزات المحتملة وضمان تمثيل جيد لكل الفئات.

- تحقيق التوازن في تصنيف البيانات:فريق متنوع قد يكون لديه فهم أعمق لتحيزات التصنيف والتفاعلات الاجتماعية. يمكن لأفراد الفريق أن يساهموا في تحقيق توازن أفضل في تصنيف البيانات وتجنب التحيزات غير المقصودة.

- استفادة من خبرات متنوعة في تصميم البيانات: يمكن لأعضاء الفريق الذين يمتلكون خلفيات مختلفة أن يساهموا في تصميم مجموعات البيانات بطرق تضمن التنوع والشمول.

- التحقق من التحيزات بشكل فعال:يمكن لأفراد الفريق أن يقوموا بتحليل وفحص البيانات بشكل أفضل للكشف عن أي تحيزات محتملة واتخاذ الإجراءات اللازمة للتصدي لها.

- تطوير حلول مبتكرة:التنوع في الفريق يمكن أن يؤدي إلى إبداع أكبر في العثور على حلول لمشاكل التحيز في نماذج التعلم الآلي. الأفراد ذوي الخبرات المتنوعة يمكنهم تقديم وجهات نظر فريدة وحلول إبداعية.

باستخدام فريق متنوع، يمكن للشركات تحسين عمليات تصنيف البيانات وتقليل التحيز في نماذج التعلم الآلي، مما يؤدي إلى نتائج أكثر عدالة وشمولاً.

1-3- الخوارزميات وحدها ليست كافية: على الرغم من أن الخوارزميات تلعب دورًا حاسمًا في تطوير الذكاء الاصطناعي، إلا أن فهم السياق والتفاعلات البشرية داخل النظام يمثل جزءًا أساسيًا من مكافحة التحيز ، بحيث تعتبر المعرفة البشرية والوعي بالسياقات الاجتماعية والثقافية جوانب مهمة لتحديد وفهم أنماط التحيز في الذكاء الاصطناعي مما يمكن لمطوري برامج الذكاء الاصطناعي أن يستفيدوا بشكل كبير من فهمهم لهذه السياقات لتصميم نماذج تكنولوجية أكثر حيادية و يسمح لهم بتحديد وتصحيح التحيزات المحتملة بشكل أفضل، وبالتالي يساهم في تطوير نماذج ذكاء اصطناعي أكثر عدالة وشمولية.

1-4- تقليل الإقصاء: يعد اختيار الميزة (Featureselection) أمرًا أساسيًا للمساعدة في تقليل الإقصاء في الذكاء الاصطناعي، وهي عملية تقوم على تقليل عدد المتغيرات المدخلة إلى نماذج الذكاء الاصطناعي بهدف تحسين أداء التنبؤ بالنتائج. تقوم هذه الخطوة باستبعاد عناصر البيانات التي لا تتمتع بتباين كافٍ للتأثير على النتائج.

2- اختبار التذوق الأعمى والذكاء الاصطناعي: زيادة على الممارسات التي ذكرناها آنفاً للحد من تحيز الذكاء الاصطناعي، يمكن أن نستند أيضاً على ما يُعرف بـ اختبار التذوق الأعمى. يعتبر هذا الاختبار وسيلة تم استخدامها لعقود للتحقق من عدم وجود تحيز. انتشر هذا الاختبار في منتصف سبعينات القرن الماضي، عندما أطلقت شركة مشروبات غازية تحدياً يتعلق بتذوق نوعين من المشروبات دون الكشف عن مصدر المشروب. تم إزالة الملصقات عن عبوات المشروب للمشاركين، وكانت نتيجة التحدي لصالح الشركة المطلقة للتحدي، حيث اختارت الأغلبية المشروب الذي كانت الشركة قد قدمته على حساب المنتج الأكثر مبيعاً، حتى على الرغم من وجود تحيز تجاري لصالح المنتج الآخر في الواقع¹⁵.

تجسد التجربة السابقة حقيقة أن إزالة معلومات التعرف عن المنتج أزلت العنصرية الرقمية، وبناءً على هذا المفهوم، يمكن للخوارزميات أن تستفيد من اختبار التذوق الأعمى لضمان عدم تأثير المعلومات المحتملة للتحيز على النتائج. يبرز هذا النهج أهمية تصميم نظم الذكاء الاصطناعي بطرق تحقق التعامل العادل والشمولي دون التأثير المباشر للمعلومات المتحيزة.

كنموذج عن ذلك، تستخدم ولاية بنسلفانيا بالولايات المتحدة الأمريكية منذ أغسطس 2016، في مقاطعة أليغيني أداة فحص تدعى (AFST) Allegheny Family Screening Tool لتحليل مئات العناصر من البيانات لكل شخص متورط في ادعاء إساءة معاملة الأطفال.

في مجال دراسة الحالات، يعتمد العاملون على بلاغات حول احتمال تعرض الأطفال للإساءة، بالإضافة إلى البيانات العامة حول العائلات المعنية. يتم تشغيل نموذج تعلم آلي يتنبأ بمستوى المخاطر من 1 إلى 20، حيث يتم فتح تحقيق إذا كانت الدرجة مرتفعة. ومع ذلك، يواجه نظام AFST تحيزات بشرية، حيث يأخذ في اعتباره المكالمات من مقدمي الرعاية الصحية إلى الخط الساخن، والتي قد تتضمن تحيزات.

بناءً على بعض الأدلة، يتضح أن المكالمات قد تظهر ارتباطاً أكبر بينها وبين العائلات ذات البشرة السمراء أو الأصل المزدوج بنسبة تصل إلى ثلاث مرات أكثر من العائلات ذات البشرة البيضاء. ورغم إقصاء العديد من هذه المكالمات، يستند النظام عليها في تعيين حدة المخاطر، مما يؤدي إلى فتح تحقيقات متحيزة رقمياً إذا كان الأشخاص المتصلون على الخط الساخن أكثر احتمالاً للإبلاغ عن العائلات ذات البشرة السمراء. في هذا السياق، يُمكن إجراء "اختبار التذوق الأعمى" من خلال تدريب النموذج على كل البيانات المتاحة لتنبؤ احتمال تعرض الأطفال للإساءة، بما في ذلك المكالمات ذات

¹⁵-مسار، مرجع سبق ذكره.

الصلة بالخط الساخن لخدمة المجتمع. بعد ذلك، يتم إعادة تدريب النموذج على كل البيانات باستثناء هذا العنصر. إذا كانت توقعات النموذج متسقة بغض النظر عن المكالمات، فإن هذا يشير إلى أن النموذج يجري توقعات بشكل عمياء. وإذا كانت التوقعات تتغير عند تضمين هذه المكالمات، فإن هذا يُظهر أن المكالمات تُعتبر متغيرًا توضيحيًا في النموذج، أو قد يكون هناك تحيز محتمل في البيانات يتعين فحصه بدقة قبل الاعتماد على الخوارزمية.

3- أدوات للتغلب على تحيز الذكاء الاصطناعي

3-1 - IBM AI Fairness 360: يتم استخدام نماذج التعلم الآلي بشكل متزايد لاتخاذ قرارات عالية المخاطر بشأن الأشخاص. على الرغم من أن التعلم الآلي، بطبيعته، هو دائمًا شكل من أشكال التمييز الإحصائي، إلا أن التمييز يصبح مرفوضًا عندما يضع مجموعات مميزة معينة في ميزة منهجية وبعض المجموعات المحرومة في وضع غير مؤات بشكل منهجي. التحيزات في بيانات التدريب، بسبب التحيز في التسميات أو نقص الإفراط في أخذ العينات، تؤدي إلى نماذج ذات تحيز غير مرغوب فيه. تتضمن حزمة IBM AI Fairness 360 Python مجموعة شاملة من المقاييس لمجموعات البيانات والنماذج لاختبار التحيزات، وتفسيرات لهذه المقاييس، والخوارزميات لتخفيف التحيز في مجموعات البيانات والنماذج. يوفر العرض التوضيحي التفاعلي IBM AI Fairness 360 مقدمة لطيفة للمفاهيم والقدرات. تقدم البرامج التعليمية ودفاتر الملاحظات الأخرى مقدمة أعمق موجهة نحو علماء البيانات. واجهة برمجة التطبيقات الكاملة متاحة أيضًا. نظرًا لكونها مجموعة شاملة من الإمكانيات، فقد يكون من المفيد معرفة المقاييس والخوارزميات الأكثر ملاءمة لحالة استخدام معينة. وللمساعدة، قامت مجموعة IBM AI FAIRNESS 360 بإنشاء بعض المواد الإرشادية التي يمكن الرجوع إليها. مع تطوير الحزمة و وضع القابلية للتوسعة في الاعتبار.

3-2 - Fairlearn: هي عبارة عن حزمة Python مقدمة من شركة "مايكروسوفت" تمكن مطوري أنظمة الذكاء الاصطناعي من تقييم عدالة نظامهم و التخفيف من أي مشكلات ملحوظة تتعلق بالتحيز الرقمي. يحتوي Fairlearn على خوارزميات التخفيف و لوحة تحكم تفاعلية¹⁶. كما يطمح المشروع ليشمل الموارد التعليمية التي تغطي العمليات التنظيمية والتقنية للحد من التحيز في الذكاء الاصطناعي (دليل مستخدم شامل، دراسات حالة مفصلة، تقارير فنية، إلخ)¹⁷.

3-3 - Fairlens: وهي مكتبة بايثون مفتوحة المصدر يمكنها الكشف عن تحيز البيانات وقياسها، من تقييم التحيز العنصري في خوارزمية COMPAS (ملف تعريف إدارة المجرمين التصحيحية للعقوبات

¹⁶-<https://pypi.org/project/fairlearn/0.4.5/>, consulté le 12 février 2024 à 13h 10 mn

¹⁷- مسار، المرجع نفسه.

البديلة). يتم استخدام هذه الخوارزمية من قبل القضاة ومسؤولي الإفراج المشروط لتقييم خطر العودة إلى الإجرام بالنسبة للأشخاص الذين لديهم إدانات جنائية. كان التقييم باستخدام FairLens مستوحى جزئيًا من تحقيق ProPublica الذي وجد أن خوارزمية COMPAS بالغت في التنبؤ بعودة الأشخاص السود ولم تتنبأ بها بشكل كافٍ بالنسبة للأشخاص البيض. منذ هذا التحقيق، استمرت الأبحاث الأخرى في العثور على مشكلات في الخوارزمية. باستخدام مجموعة بيانات COMPAS ، والتي تتضمن نتائج خوارزمية COMPAS ، حددت FairLens المتغيرات المرتبطة بالسماوات الحساسة. تتوافق هذه المتغيرات مع الجنس والعمر والعرق. وقد قدم استخدام FairLens لتصوير تلك السماوات حسب درجة المخاطر صورة واضحة مفادها أن التباين الأكبر بين الدرجات كان في الواقع بسبب العرق. بمعنى آخر، مع زيادة درجة المخاطرة، زاد أيضًا عدد الأشخاص السود، بينما انخفض عدد الأشخاص البيض. ثم قدمت مقاييس FairLens قياسًا أكثر دقة للمسافة الإحصائية لتوزيع درجات المخاطر في كل مجموعة فرعية وفي مجموعة البيانات بأكملها. يمكن بعد ذلك لمسجل عدالة FairLens استخدام اختبارات الفرضيات ذات الصلة لتحديد الاختلافات في توزيعات درجات المخاطر حسب كل مجموعة فرعية، مثل العرق. أخيرًا، تمكنت شركة FairLens من العثور على أنه حتى عند إسقاط العرق كسمة، تظل توقعات العودة إلى الإجرام متحيزة تجاه الأشخاص من مجموعات عرقية مختلفة¹⁸.

3-4-TCAV: هو نظام أُعلن عنه خلال مؤتمر «TCAV» عام 2019 من قبل سونداربيتشاي،

الرئيس التنفيذي لشركة Google. يُعد مبادرة بحثية باسم "Test With Concept Activation (TCAV)"، وهو آلية تفتيش تستخدم للكشف عن أنماط التحيز الرقمي في أنظمة الذكاء الاصطناعي. يمكن للنظام فحص النماذج لتحديد العناصر التي يمكن أن تتسبب في التحيز بناءً على العديد من العوامل مثل العرق والدخل والموقع وغيرها. يتعلم نظام "TCAV المفاهيم" من خلال الأمثلة، مما يساعد في فهم وتحليل البيانات بشكل أفضل.

"Google What-If Tool" هي أداة أنشئت من قبل باحثين ومصممين في Google لتكون موردًا عمليًا لمطوري أنظمة التعلم الآلي. تُحاول هذه الأداة الإجابة على سؤال صعب حول "ما هو الإنصاف الذي يريده المستخدمون؟" من خلال توفير واجهة تفاعلية مفتوحة المصدر تمكن المستخدمين من استكشاف نماذج التعلم الآلي بصورة بصرية. تُمكن أدوات What-If TensorBoard مفتوحة المصدر، من تحليل مجموعات البيانات وفهم كيفية عمل نماذج التعلم الآلي في مختلف السيناريوهات.

¹⁸-OECD.AI, Policy Observatory, FairLens detected racial bias in a recidivism prediction algorithm, publié le 19 avril 2023, <https://oecd.ai/en/catalogue/tools/fairlens/tool-use-cases/fairlens-detected-racial-bias-in-a-recidivism-prediction-algorithm>, consulté le 15 février 2024 à 11h 50mn

بفضل واجهة المستخدم الرسومية والبسيطة، يُسهّل استخدام أداة What-If على المستخدمين، حتى إن كانوا غير مبرمجين، استكشاف واختبار نماذج التعلم الآلي وتحليل أدائها.

Skater -5-3: مبادرة أطلقتها Oracle ، وهي عبارة عن مكتبة Python لإزالة الغموض عن نموذج الصندوق الأسود، وهي نماذج تنشأ مباشرة من البيانات وبواسطة الخوارزميات، وبالتالي لا يمكن معرفة الكيفية التي يتم بها دمج المتغيرات لصنع التنبؤات. يساعد Skater على بناء نظام تعلم آلي يساهم في زيادة الثقة في النموذج وتوفير الشفافية في عمليات اتخاذ القرار¹⁹

الخاتمة:تكنولوجيا الذكاء الاصطناعي، كغيرها من التكنولوجيات الحديثة، تأتي مع فوائد هائلة للبشرية، ولكنها تحمل أيضًا جوانب خطيرة ومظلمة. يظهر ذلك بشكل خاص عندما يتعلق الأمر بحقوق الأفراد وحررياتهم، حيث تزيد خطورة الأمور عند وقوع الأخطاء.

تتوعد أسباب تلك التجاوزات في أنظمة الذكاء الاصطناعي، وكان من بينها التحيز الرقمي الذي يمكن أن ينشأ من مصادر متعددة. تكمن أهمية الجهود المبذولة التي تسعى لتطوير آليات وحلول للحد من هذا التحيز أهمية في المحافظة على التوازن بين التقدم التكنولوجي وحقوق الفرد وحرياته. علما أن هذا خلصت الدراسة الى جملة من النتائج في مجملها كالآتي:

❖ التحيز الرقمي او العنصرية الرقمية كما عرفه العلماء هو حالة انحراف في نتائج خوارزميات التعلم الآلي يحدث بسبب وجود فرضيات مُتحيّزة أثناء عملية تطوير الخوارزمية، وتمثل انعكاسًا لعنصرية المجتمع وتحيزه ضد فئة معينة، أو قد يكون نتيجة تحيز في بيانات التدريب التي يتم تغذية نظام الذكاء الاصطناعي بها.

❖ يظهر التمييز الرقمي كتمديد للتمييز البشري، حيث يتجلى هذا التمييز بشكل عنصري تجاه بعض الفئات. يمكن رؤية ذلك بوضوح في أنظمة التعرف على الوجوه، حيث تتميز بقدرتها على التعرف على البشرة ذات اللون الفاتح أكثر من البشرة ذات اللون الداكن، وهو أمر يسلط الضوء على التمييز العنصري. وفي سياق الأمان، يُلاحظ أيضاً انحيازاً في استخدام تكنولوجيا التمييز الرقمي حيث يتم توجيه الاتهامات بشكل غير عادل نحو الفئات ذات البشرة الداكنة، مما يجعلهم ضحية لهذا التمييز.

❖ تتعدد الاسباب التي تدفع بالخوارزميات للتحيز منها إذا كانت البيانات المستخدمة للتدريب تحتوي على تحيز، سواء كان ذلك نتيجة لتمثيل غير كافي لفئة معينة أو تحديد خاطئ للعوامل المؤثرة، فإن الخوارزمية قد تعكس هذا التحيز في النتائج.

¹⁹مسار، المرجع نفسه.

العنصرية الرقمية في أنظمة الذكاء الاصطناعي

❖ إدارة التحيز الخوارزمي تتطلب مجموعة من الآليات والممارسات، ومن بينها فحص البنات وتنويع مصادرها تحسين التنوع في فرق العمل، تطبيق مبادئ الشفافية التدريب والوعي، تقييم دوري للخوارزميات.

❖ الالتزام بأخلاقيات الذكاء الاصطناعي مثل ما أوصت به المنظمات الدولية كالينسكو والاتحاد الأوروبي من أجل تعزيز وحماية حقوق الإنسان والكرامة الإنسانية والاستدامة البيئية.

التوصيات

■ تعزيز الوعي بين المجتمعات حول مخاطر التحيز والعنصرية في نظم الذكاء الاصطناعي. يمكن ذلك عبر حملات توعية، وورش عمل، ومشاركة معلومات حول كيفية تأثير هذه القضايا في الحياة الواقعية والرقمية.

■ ضرورة وجود قوانين داخلية في الدول التي تتبنى التكنولوجيا الذكية. يجب أن تكون هذه القوانين تحتوي على نصوص وقواعد تحظر إنتاج أنظمة ذكاء اصطناعي تتسم بالتحيز والعنصرية، مع فرض عقوبات على المخترقين لهذه القواعد.

■ ضرورة وجود اتفاقيات دولية خاصة تتعامل بشكل صارم مع مسألة التحيز والعنصرية في نظم الذكاء الاصطناعي. يجب أن تحدد هذه الاتفاقيات بنودًا صارمة تعاقب الشركات المنتجة لتلك الأنظمة في حالة انتهاكها للمعايير الأخلاقية وحقوق الإنسان.

■ تشجيع الشركات على نشر معلومات شفافة حول كيفية تدريب وعمل النماذج الذكية، وتحفيز المساءلة في حالة وجود تحيز أو سلوك غير أخلاقي.

■ إقامة حوار مستمر مع المجتمعات المحلية والعالمية لفهم احتياجاتها ومخاوفها حيال تقنيات الذكاء الاصطناعي، وتكامل هذا الفهم في تطوير وتحسين النظم.