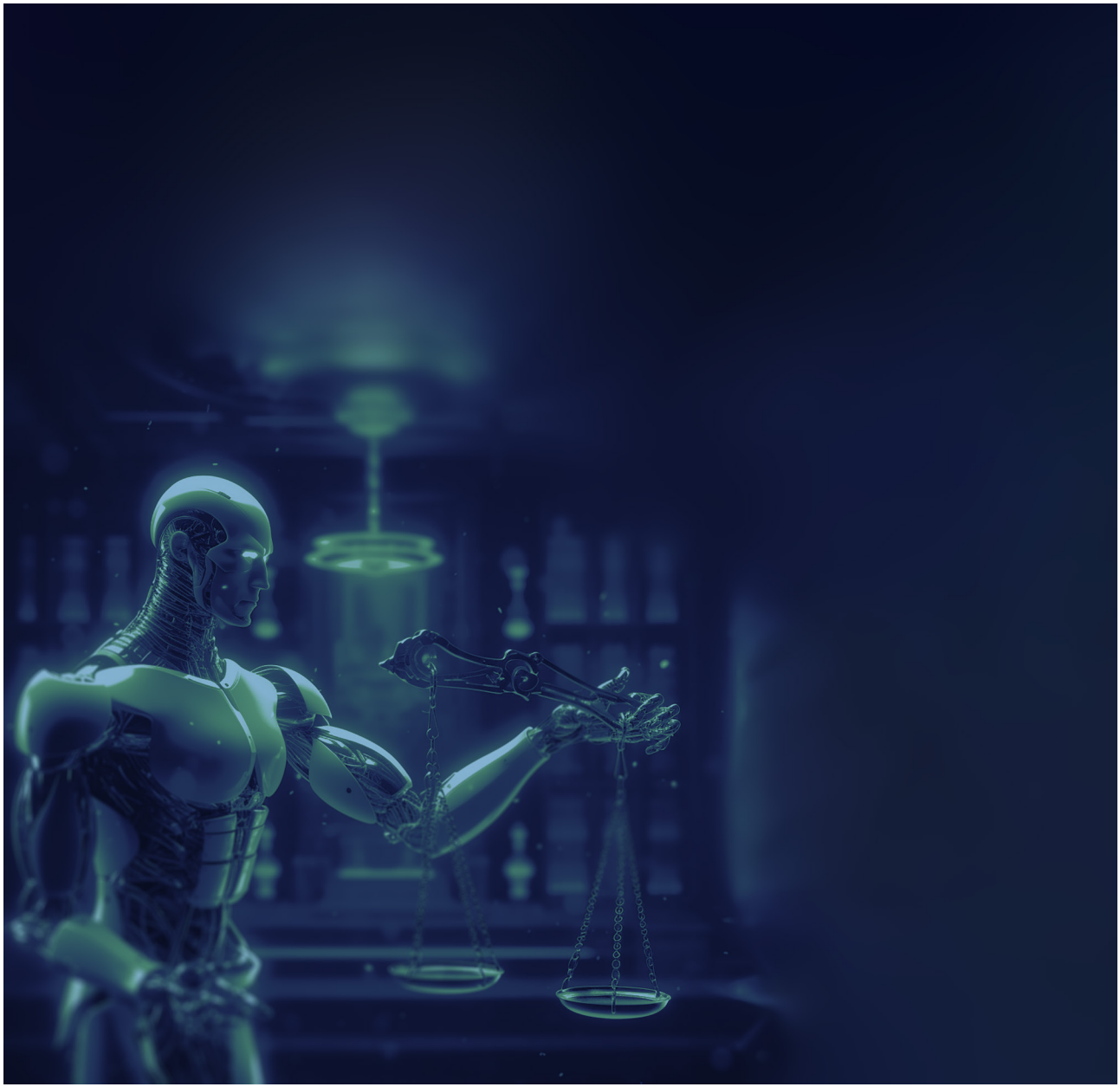
SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

التحيز في أنظمة الذكاء الاصطناعي

تحديات وحلول

أغسطس 2024

نـــوع الوثيقـة | دراسات وتقارير
تصنيف الوثيقـة | عام
رقـــم الوثيقـة | SDAIA-P101
رقـــم الإصـدار | 1.0



الملخص التنفيذي

توسعت كثير من الدول والمنظمات والشركات العالمية في تفعيل تطبيقات الذكاء الاصطناعي بشكل مضطرد في عدد من القطاعات الحساسة والحيوية، بما في ذلك القطاعات العدلية والرعاية الصحية والتعليم. وترافق مع هذا الاهتمام والتوسع قلقاً متنامياً من تسرب التحيز إلى أنظمة الذكاء الاصطناعي، لا سيما تلك المرتبطة باتخاذ القرار. ولا شك أن مرور أنظمة الذكاء الاصطناعي بالعديد من المراحل والعمليات وانخراط العديد من العاملين في هذا النوع من المشاريع على اختلاف مهامهم وتخصصاتهم وثقافاتهم، يزيد من مخاطر وجود التحيزات خلال مراحل البناء أو التطوير. وهذه التحيزات قد تقوض فائدة أنظمة الذكاء الاصطناعي؛ لأن الذكاء الاصطناعي قد يُصبح أداة لتضخيم التحيزات وزيادتها بدلاً من تلافيها، وقد يسيء لسمعة المنظمة، ناهيك عن المسؤوليات القانونية المرتبطة بنتائج وآثار تلك الأنظمة.

وبشكل عام، يحدث التحيز في أنظمة الذكاء الاصطناعي لسببين رئيسين:

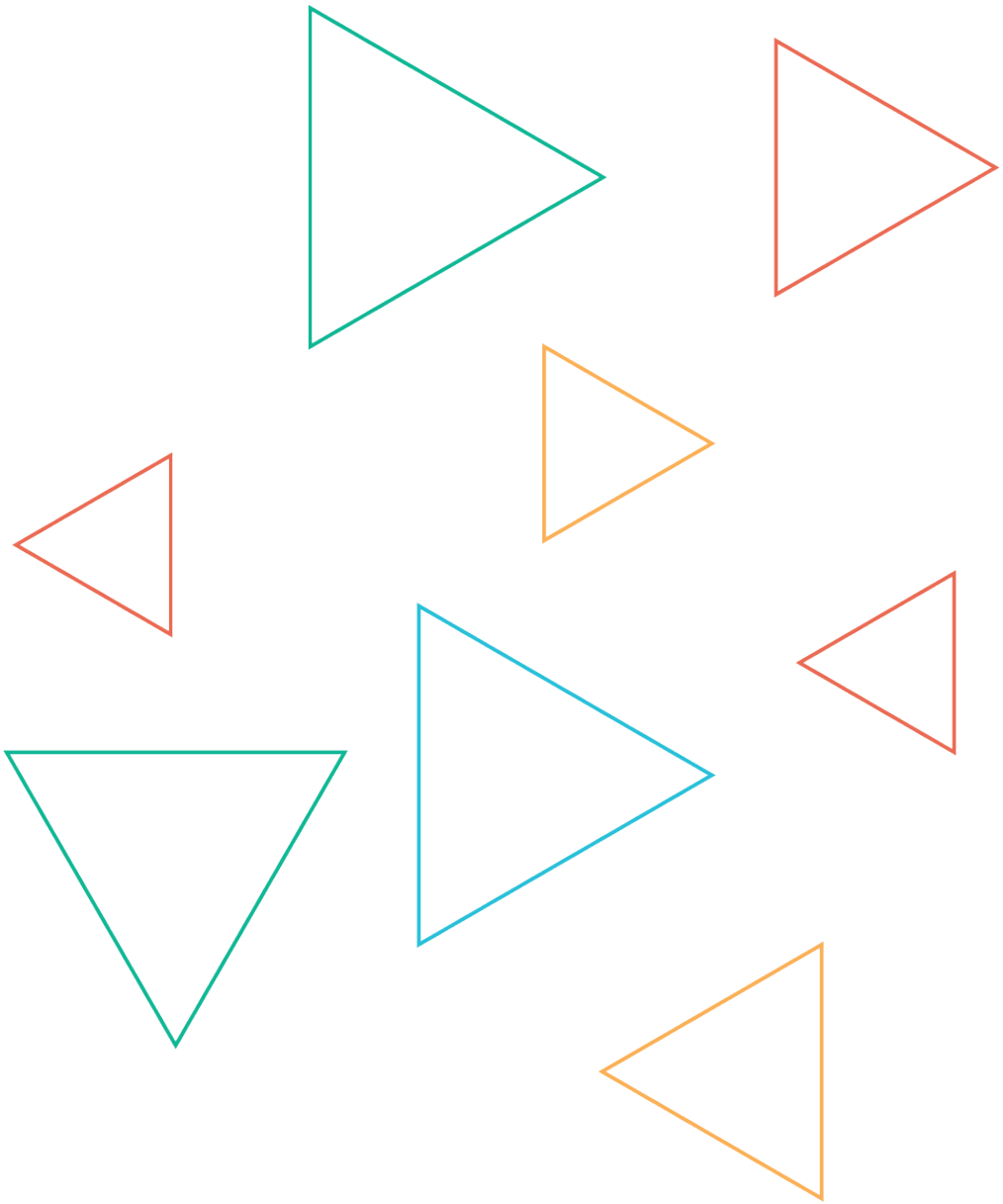
- التحيزات التقنية: وهي التحيزات التي تنشأ عن تصميم وبناء أنظمة الذكاء الاصطناعي وتطويرها وتنفيذها، والتي غالباً ما تنبع من الجوانب الفنية للتقنية نفسها، مثل التحيز في البيانات، والخوارزميات والتي قد تؤثر على عدالة أو دقة أنظمة الذكاء الاصطناعي.
- التحيزات الإدراكية: وهي التحيزات التي تؤدي إلى الانحراف عن التفكير العقلاني أو السلوك الطبيعي في اتخاذ القرار أو الحكم على الأشياء، وغالباً ما تنبع هذه التحيزات من التصورات الذهنية أو الاختصارات العقلية التي يستخدمها العقل البشري لتبسيط وتسريع عمليات الفهم واتخاذ القرارات.

وقد حدد علماء النفس أكثر من 180 تحيزاً بشرياً وصنفوها إلى عدة أنواع. ويمكن أن تتسرب التحيزات إلى خوارزميات تعلم الآلة عن غير قصد عن طريق مجموعة بيانات التدريب التي تتضمن تلك التحيزات.

تهدف هذه الدراسة إلى توضيح جذور التحيزات الأكثر شيوعاً التي يمكن أن تقع خلال المراحل المختلفة لدورة حياة بناء أنظمة الذكاء الاصطناعي، وتتضمن: تعريف المشكلة، وتجهيز البيانات، وبناء النموذج، واختبار النموذج، ونشر النموذج. وفي حين أنه لا يوجد حل شامل وكامل للتخلص من جميع التحيزات على اختلاف أنواعها، إلا أن هناك عدد من الطرق والأساليب والأدوات والممارسات التي تسعى إلى إدارة آثار التحيزات على مخرجات الخوارزميات والتقليل منها، وتسهم في اكتشافها مبكراً ومراقبتها والتحكم فيها. كما تتضمن الدراسة إرشادات عامة مهمة لتكوين ثقافة في بيئة العمل لتهيئة الموظفين لاكتشاف التحيزات، والإبلاغ عنها، والتعامل معها، فضلاً عن إرشادات خاصة بكيفية التعامل مع التحيزات التي تنشأ في كل مرحلة من مراحل بناء أنظمة الذكاء الاصطناعي.

وتتمثل إحدى النتائج الرئيسية لهذه الدراسة في أن أنظمة الذكاء الاصطناعي يمكن أن تكون متحيزة بعدة طرق، وغالباً ما تكون هذه التحيزات غير مقصودة. إذ من الممكن أن تنشأ التحيزات من البيانات المستخدمة لتدريب الأنظمة، أو الخوارزميات المستخدمة لمعالجة تلك البيانات، أو حتى طريقة عرض النتائج على المستخدمين. ويمكن أن تؤدي هذه التحيزات إلى التمييز ضد فئات معينة من الناس، وتكريس الصور النمطية، وترسيخ عدم المساواة.

ولذلك لابد من مراقبة ومراجعة النماذج قيد التشغيل، وجعل هذه العملية مستمرة، لضمان الالتزام بالسياسات الأخلاقية المعتمدة، مما يساعد في الكشف عن التحيزات مبكراً وتخفيفها قبل استفحال ضررها. والبقاء على اطلاع بالبحوث الصادرة في هذا المجال، واختيار الموردين الملتزمين بتطوير تطبيقات الذكاء الاصطناعي المسؤول



المحتويات

1. مقدمة

تتسابق العديد من الدول والمنظمات للاستفادة من قدرات الذكاء الاصطناعي في معالجة البيانات الضخمة والوصول إلى نتائج ذات قيمة في رفع الكفاءة ودعم اتخاذ القرار. وتأتي تطبيقات اتخاذ القرار والتنبؤ بالمستقبل على رأس الأولويات لغرض استبعاد تأثير العنصر البشري، لا سيما فيما يتعلق بالعدالة والحياد والإنصاف. إلا أن الأمر ليس بهذه السهولة، فخوارزميات الذكاء الاصطناعي ليست محصنة ضد التحيزات. وإحدى التحديات التي يواجهها علماء البيانات هي التأكد من أن البيانات التي تُدرَّب عليها خوارزميات تعلم الآلة نظيفة ودقيقة ومُصنفة جيداً – في حالة تعلم الآلة الموجه – وأيضاً يجب أن تكون خالية من أي بيانات متحيزة قد تتسبب في انحراف نتائج أنظمة الذكاء الاصطناعي. وعوضاً عن أن يكون الذكاء الاصطناعي رافداً للعدالة ومتفوقاً على القرار البشري، قد يصبح مساهماً في نشر التحيز وعدم الإنصاف.

ويشير تقرير صادر عن كلية هارفارد للأعمال (Harvard Business School) وشركة أكسنتشر (Accenture) إلى أن ما يقرب من 27 مليون عامل في الولايات المتحدة قد يستعبدون من الترشح للعمل؛ إذ إن الشركات التي تعتمد على الذكاء الاصطناعي – في بعض الأحيان – ترفض المرشحين المحتملين مما يعني أن هؤلاء الأشخاص "مخفيين" عن جهات التوظيف[1].

ويظهر تحيز أنظمة الذكاء الاصطناعي عندما تتخذ قرارات غير عادلة بشكل منهجي لمجموعات معينة من الناس[2]. ويعد تحيز الذكاء الاصطناعي حالة شاذة في مخرجات خوارزميات تعلم الآلة، بسبب الافتراضات المتحيزة أثناء عملية بناء الخوارزميات أو الأحكام المسبقة في بيانات التدريب، ومع أن المفترض أن تكون التقنيات محايدة، ولا تنطوي على أي نوع من أنواع التحيزات، إلا إن أنظمة الذكاء الاصطناعي تعكس تحيزات الأشخاص الذين يعملون بها على اختلاف مراحل بناء هذه الأنظمة. وقد تساعد أنظمة الذكاء الاصطناعي على تضخيم الآثار السلبية لهذه التحيزات. ومما يزيد الأمر سوءاً، أنه قد يكون من الصعب اكتشاف هذه التحيزات قبل البدء في عملية التنفيذ. ويمكن أن تشمل النتائج السلبية المحتملة خسارة الأرباح والضرر بالسمعة والتعرض لشكاوى المستهلكين والمسؤوليات القانونية؛ ولذلك يُعد فهم جذور التحيزات في دورة حياة بناء أنظمة الذكاء الاصطناعي هو الخطوة الأولى لحلها وتحسين معدلات نجاح مشاريع الذكاء الاصطناعي، إذ إن وجود التحيزات يعد من أهم أسباب فشل مشاريع الذكاء الاصطناعي، لا سيما في مجال اتخاذ القرار.

ويمكن أن يدخل الخطأ البشري والتحيز وسوء التقدير في دورة حياة بناء أنظمة الذكاء الاصطناعي ويؤدي إلى خلق تحيزات في تلك الأنظمة بصورة مقصودة أو غير مقصودة. وهذه التحيزات قد تقع في المراحل الأولية لصياغة المشكلة وجمع البيانات واستكشافها إلى المراحل الحرجة كبناء النموذج والتنفيذ. وفي حين أنه قد يكون من المتعذر إزالة التحيزات تماماً من تطبيقات الذكاء الاصطناعي، إلا إن فهم مصادرها وكيفية ظهورها ووجود سياسات وإجراءات وضوابط مدروسة وأدوات تقنية مساعدة يمكن أن يساعد في تجنب كثير منها ويخفف من حدة آثارها، وخاصة عند وجود مجموعات بيانات متوازنة ومتنوعة وبرامج حوكمة متطورة، وتدريب الفرق العاملة في مشاريع الذكاء الاصطناعي على معرفة كيفية تطبيق الخوارزميات بشكل صحيح واختبار النتائج باستمرار بحثاً عن الانحرافات المحتملة.

تسعى هذه الدراسة إلى توضيح جذور أبرز أنواع التحيزات التي يمكن أن تقع خلال دورة حياة بناء أنظمة الذكاء الاصطناعي، وتصنيفها حسب كل مرحلة من مراحل دورة الحياة، مما يساعد على تجنب حدوث هذه التحيزات أو تقليلها، والتطرق إلى أفضل الطرق والممارسات والتجارب العالمية في اكتشافها مبكراً والتعامل معها، وتسليط الضوء على عدد من المبادرات والأدوات التي أصدرتها المنظمات الدولية المهتمة بأخلاقيات الذكاء الاصطناعي.

2. دورة حياة بناء أنظمة الذكاء الاصطناعي

يساعد فهم دورة حياة بناء أنظمة الذكاء الاصطناعي على تحديد التفاصيل والمراحل التي تحتاج إلى العناية والاهتمام لمعالجة مصادر التحيز، مثل مزيد من المعلومات حول المخرجات والأدوار الوظيفية التي يجب تعيينها في كل مرحلة مما يساهم في تحديد المسؤوليات وتوزيع المهام، كما يساعد على تحديد المخاطر التي يمكن أن تقع في كل مرحلة ودرجة خطورتها.

كثير من أنظمة الذكاء الاصطناعي الموجودة اليوم تعتمد على تقنيات تعلم الآلة، بمعنى أن الآلة تتعلم من البيانات عن طريق تحليلها وفهم العلاقات والأنماط الموجودة فيها بهدف تنفيذ مهام محددة أو دعم اتخاذ القرار أو التنبؤ بالمستقبل. وبصورة عامة، يمكن تقسيم مراحل بناء هذه الأنظمة إلى أربع مراحل موضحة في **الشكل (1):** تعريف المشكلة، وتجهيز البيانات، وبناء النموذج، واختبار النموذج، ونشر النموذج.

الشكل (1): دورة حياة بناء أنظمة الذكاء الاصطناعي (تعلم الآلة)

2.1 تعريف المشكلة

هي عملية تعريف المهمة التي يمكن حلها أو أتمتتها باستخدام تقنيات الذكاء الاصطناعي. إن تحديد الهدف، وتعريف المشكلة، ووضع حدود ونطاق المشكلة، و تحديد القيمة التي ستنعكس على المستفيدين النهائيين من هذا المشروع، هي من المراحل الأولية و الأساسية لمشروع الذكاء الاصطناعي.

2.1.1 تعريف المشكلة

تعريف المشكلة (Problem Identification) الُمراد فهم المشكلة بشكل واضح ووضع حل لها واكتشاف العوامل المختلفة التي تؤثر عليها وتحديد الهدف من حلها، مع التركيز على قيود العمل قبل الشروع في إيجاد الحل.

2.1.2 تحديد نطاق المشكلة

عملية تحديد نطاق المشكلة (Problem Scoping) تهدف إلى تعريف حدود المشكلة مما يسهم في التركيز على إيجاد الحل المناسب. وللمساعدة في تحديد نطاق المشكلة يمكن استخدام طريقة طرح أربعة أسئلة : من؟، وماذا؟، وأين؟، ولماذا؟.

"من؟": معرفة الأشخاص المتأثرين بشكل مباشر وغير مباشر بالمشكلة (المعنيين).

"ماذا؟": تحديد طبيعة المشكلة وجمع الأدلة لإثبات أن المشكلة التي تم تعريفها موجودة بالفعل.

"أين؟": معرفة أين تنشأ المشكلة ومواضعها.

"لماذا؟": السبب لحل المشكلة وما هي الفوائد التي تعود على المعنيين بعد حل المشكلة.

2.2 تجهيز البيانات

2.2.1 جمع البيانات

جمع البيانات (Data Acquisition) اللازمة لحل المشكلة التي تم تعريفها وتحديد نطاقها في المرحلة السابقة. وعملية جمع البيانات يجب أن تكون دقيقة وشفافة ومن مصادر موثوقة. ويمكن أن تكون البيانات بتنسيقات متنوعة، مثل: نصوص وفيديوهات وصور وأصوات، ويمكن جمعها من مصادر مختلفة، مثل: قواعد البيانات وإجراء المسوح وصفحات شبكة الإنترنت والمجلات والصحف والسجلات والكاميرات وأجهزة الاستشعار. وتُعد مرحلة جمع البيانات من أكثر المراحل عرضة لحدوث التحيزات.

2.2.2 تجهيز البيانات

تجهيز البيانات (Data Preparation) هي عملية ترتيب البيانات وتنظيفها وتجهيزها لاختيار الخوارزمية المناسبة وبناء النموذج المطلوب. ويمكن ترتيب البيانات في شكل جداول أو رسومات أو مخططات أو قواعد بيانات. ويتضمن ذلك استكشاف البيانات (Data Exploring) لمعرفة الميزات الأكثر أهمية في مجموعة البيانات وما هو الاتجاه العام لهذه البيانات. ويعد استكشاف البيانات الخطوة الأولى في تحليل البيانات وتصويرها للكشف عن طبيعة البيانات وفهم أنماطها.

2.3 بناء النموذج

بعد مرحلة جمع البيانات وتجهيزها واستكشافها، تأتي مرحلة بناء النموذج (Model Development) أو ما يُسمى بالنمذجة (Modelling) التي يمكن عن طريقها صياغة علاقات رياضية تقريبية لتمثيل البيانات. وتُعد القدرة على وصف العلاقات بين المتغيرات رياضياً هي جوهر كل نموذج ذكاء اصطناعي، ويحتاج كل نموذج إلى نهج رياضي مناسب لتحليل البيانات.

وبصورة عامة، هناك نهجان مشهوران لنمذجة الذكاء الاصطناعي:

2.3.1 النهج القائم على القواعد

يعتمد النهج القائم على القواعد بشكل عام على القواعد المحددة مسبقاً والتي تحدد كيفية اتخاذ القرارات أو الإجراءات بناء على بيانات الإدخال. على سبيل المثال، يمكن استخدام نظام قائم على القواعد لتشخيص الحالات الطبية بناءً على الأعراض. وعلى الرغم من أن النهج القائم على القواعد يعتبر من أقدم وأبسط الطرق المستخدمة في الذكاء الاصطناعي، إلا أنها لا تزال تستخدم في العديد من التطبيقات حتى اليوم، خاصةً في المجالات التي تتطلب الدقة والتحكم في العمليات المعقدة.

2.3.2 النهج القائم على التعلم

النهج القائم على التعلم هو نهج يعتمد على تدريب الأنظمة الحاسوبية على التعلم من البيانات وتحليلها واستخلاص الأنماط والمعلومات منها، وذلك بدلاً من برمجة الأنظمة بشكل يدوي. ويتم ذلك عن طريق استخدام الخوارزميات والنماذج الرياضية التي تمكن الحاسوب من التعلم والتكيف مع المهام المختلفة. ويتم تحسين هذا النهج باستمرار من خلال تحسين الخوارزميات وتوسيع قاعدة البيانات المستخدمة في التدريب. ويعد هذا النهج أحد أهم الأساليب المستخدمة في تطوير تطبيقات الذكاء الاصطناعي في مجالات مختلفة مثل تعلم الآلة والتحليل الضخم للبيانات والتعرف على الصوت والصورة واللغة الطبيعية وغيرها.

2.4 اختبار النموذج

في مرحلة اختبار النموذج (Model Testing) يتم تقييم أداء النموذج على مجموعة من البيانات الجديدة التي لم يتعرف عليها النموذج أثناء التدريب. حيث أن الغرض من هذه المرحلة هو التأكد من أن النموذج لم يقم ببساطة بحفظ بيانات التدريب ولكنه تعلم بالفعل عمل تنبؤات دقيقة بناءً على الأنماط العامة في البيانات.

وبشكل عام تنقسم مجموعة البيانات إلى ثلاثة أقسام: مجموعة بيانات التدريب (Training Data Set)، ومجموعة بيانات التحقق من الصحة (Validation Data Set) ومجموعة بيانات الاختبار (Test Data Set). وهذه المجموعات يجب ألا يكون بينها أي تشارك في البيانات. حيث إن المجموعتين الأخريتين للبيانات تستخدمان في مرحلة اختبار النموذج. تُستخدم مجموعة بيانات الاختبار لتقييم الأداء النهائي للنموذج، بينما تُستخدم مجموعة بيانات التحقق من الصحة لضبط معاملات النموذج أثناء عملية التدريب. كلا مجموعتي البيانات مهمتان لتقييم وتحسين أداء نموذج الذكاء الاصطناعي.

2.4.1 التحقق

تأتي مرحلة التحقق من صحة النموذج (Model Validation) بعد تدريب النموذج للتأكد من دقة النموذج باستخدام بيانات مجموعة التحقق، وضبط المعاملات (Parameters) للحصول على أفضل النتائج.

2.4.2 الاختبار

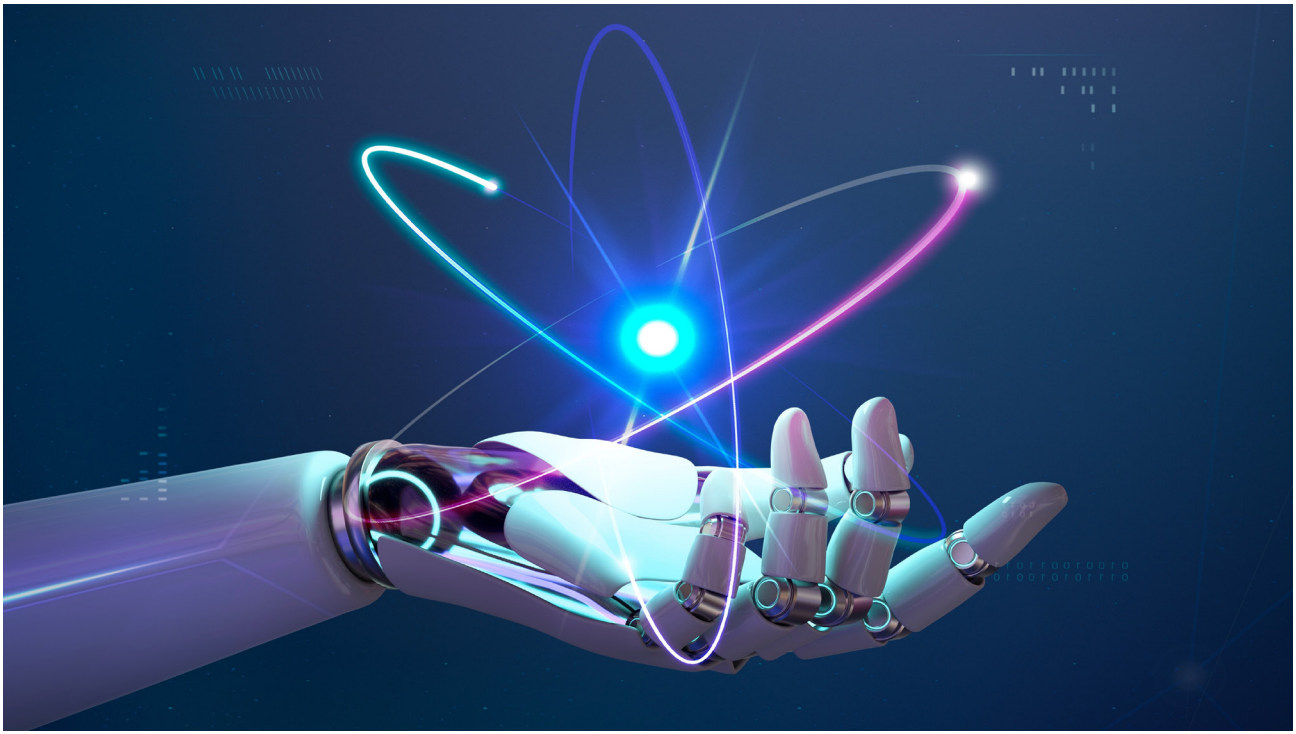
تأتي مرحلة اختبار النموذج (Model Test) بمثابة اختبار نهائي للنموذج إذ إن الخوارزمية تُختبر على مجموعة بيانات الاختبار التي

تُعد جديدة بالنسبة لها. وذلك عن طريق تغذية مجموعة بيانات الاختبار في النموذج ومقارنتها بالإجابات الفعلية. والغرض من هذا التقييم هو تقديم تقدير رياضي لمدى بعدنا عن إجراء تنبؤات صحيحة.

2.5 نشر النموذج

تُعد مرحلة نشر نماذج الذكاء الاصطناعي هي المرحلة الأخيرة في دورة بناء أنظمة الذكاء الاصطناعي، حيث يُدمج النموذج المدرب في البيئة التقنية التشغيلية للاستفادة من قدراته عن طريق المستخدمين، وتتطلب هذه المرحلة خطوات دقيقة لضمان سير العملية بسلاسة وتحقيق النتائج المرجوة، وذلك على النحو التالي:

- التكامل: دمج النموذج مع الأنظمة والتطبيقات الموجودة، مع مراعاة معايير الأمن والخصوصية.
- المراقبة: المراقبة المستمرة لأداء نماذج الذكاء الاصطناعي لرصد أي مشكلات محتملة.
- التحديث: تحديث النماذج وتحسينها بناءً على البيانات الجديدة والخبرات المكتسبة.
- الحوكمة: وضع سياسات وإجراءات لضمان استخدام نماذج الذكاء الاصطناعي بشكل مسؤول وأخلاقي.



3. مصادر التحيزات

أحد أبرز التحديات في إدارة التحيزات في أنظمة الذكاء الاصطناعي، هو معرفة وتحديد مصادر هذه التحيزات، لاحتوائها ومعالجة آثارها الضارة، فالمحاولات الحالية تركز بشكل رئيسي على العوامل الحسابية مثل عدالة الخوارميات ذاتها، بينما تغفل العوامل البشرية والاجتماعية[3]؛ ولذلك التركيز على كل مرحلة من مراحل دورة حياة بناء أنظمة الذكاء الاصطناعي يسهم بشكل كبير في فهم جذور التحيزات بصورة أفضل، والتعرف على أنواعها وكيفية حدوثها، مما يزيد من فاعلية اكتشافها وتقليل مخاطرها. يعرض **الشكل (2)** أبرز أنواع التحيزات في مراحل دورة حياة بناء أنظمة الذكاء الاصطناعي.

الشكل (2): أبرز أنواع التحيزات في مراحل دورة حياة بناء أنظمة الذكاء الاصطناعي.

تعريف المشكلة	تجهيز البيانات	بناء النموذج	اختبار النموذج	نشر النموذج
تحيز تعريف المشكلة	تحيز اختيار العينة	تحيز الخوارزمية	تحيز تسرب البيانات	تحيز انجراف البيانات
تحيز تأثير الإطار	تحيز القياس	تحيز التجميع	تحيز التقييم البشري	تحيز حلقة التغذية الراجعة
تحيز التصميم	تحيز التسمية	تحيز المتغير المحذوف	تحيز المراقب	التحيز البيئي
تحيز التفرقة العمرية	تحيز البقاء	تحيز فرط التخصيص	تحيز السبب والنتيجة	
	تحيز الإرباك	تحيز فرط التعميم		
		تحيز التمويل		

3.1 التحيزات المرتبطة بتعريف المشكلة

3.1.1 تحيز تعريف المشكلة

يمكن أن تنشأ التحيزات بناءً على كيفية تعريف المشكلة، على سبيل المثال قد ترغب شركة بطاقات الائتمان في توقع الجدارة الائتمانية للعميل، ولكن "الجدارة الائتمانية" مفهوم غامض إلى حد ما، ومن أجل ترجمته إلى شيء يمكن حسابه، يجب تحديد الهدف من حل هذه المشكلة، مثلاً: زيادة هوامش الربح أو زيادة عدد القروض التي يتم سدادها، وبعد ذلك يمكن تحديد الجدارة الائتمانية في سياق هذا الهدف[4].

3.1.2 تحيز تأثير الإطار

تأثير الإطار (Framing Effect) هو من التحيزات المعرفية في علم النفس، ويؤثر على اتخاذ قراراتنا اليومية عندما يقال بطرق مختلفة. بمعنى آخر، هناك ميل للتأثر بالطريقة التي تُقدم بها المعلومات أكثر من المحتوى نفسه. على سبيل المثال، عندما يكون هناك منتجين غذائيين مكتوب على أحدهما "10% دهون" بينما الآخر مكتوب عليه "90% خالي من الدهون"، سيؤدي تأثير التأطير إلى اختيار الخيار الثاني، إذ يبدو أنه هو الخيار الأكثر صحياً. وكمثال آخر، يفضل مندوب المبيعات الذي يروج لمنتج ما بقول أن "80% من العملاء راضون عن المنتج" بدلاً من الاعتراف بأن "20% من العملاء غير راضين"، مع العلم أن الإحصائيتين تنقلان نفس المعلومات في الواقع.

3.1.3 تحيز التصميم

هناك ارتباط وثيق بين تحيز التصميم (Design Bias) و تحديد المشكلة في سياق الذكاء الاصطناعي، لأنه يمكن أن يؤثر على كيفية تأطير المشكلة وكيفية تطوير الحل. يحدث تحيز التصميم (Design Bias) عندما تنتج الخوارزمية نتائج متحيزة بشكل منهجي بسبب افتراضات خاطئة بنيت عليها عملية تعلم الآلة. وفي بعض الأحيان، تحدث التحيزات نتيجة للقيود على النظام مثل القدرة الحاسوبية، وبناء على ذلك قد تتغير الدقة المستهدفة للمخرجات.

3.1.3.1 تحيز الترتيب

تحيز الترتيب (Ranking Bias)، هو ميل محركات البحث أو خوارزميات التوصية أو أنظمة الترتيب الأخرى إلى تفضيل أنواع معينة من النتائج على أنواع أخرى، مما يؤدي إلى تمثيل متحيز أو منحرف للمعلومات. يمكن أن يحدث هذا التحيز بسبب مجموعة متنوعة من العوامل، بما في ذلك تصميم الخوارزمية والبيانات المستخدمة لتدريب الخوارزمية وتفضيلات المستخدمين الذين يتفاعلون مع النظام. فعلى سبيل المثال يمكن فهم محرك البحث الذي يعرض ثلاث نتائج لكل شاشة أن الغرض من ذلك هو تمييز النتائج الثلاثة الأولى أكثر قليلاً من النتائج الثلاثة التالية.

3.1.3.2 تحيز العرض

تحيز العرض (Presentation Bias) مشتق من حقيقة أنه لا يمكنك تلقي تعليقات المستخدمين إلا على العناصر التي تم تقديمها إلى المستخدم، ويتأثر احتمال تلقي تعليقات المستخدم بشكل أكبر بمكان عرض العنصر[5].

3.1.4 تحيز التفرقة العمرية

التفرقة العمرية هي تحيز مجتمعي تجاه كبار السن يظهر في بعض السياسات والممارسات المجتمعية والمؤسسية. وتركز معظم الأبحاث الحالية في مجال التحيزات في الذكاء الاصطناعي إلى حد كبير على تحيز العرق والجنس والعواقب الوخيمة التي تنشأ نتيجة لذلك؛ بينما لا تولي كثير من تلك الأبحاث اهتماماً بالتحيز المرتبط بالتفرقة العمرية (Ageism Bias)، وهذا قد يؤدي إلى استبعاد كبار السن من تطور التقنية و تفاقم الفجوة الرقمية مع كبار السن[6].

ويدعو تقرير صادر عن منظمة الصحة العالمية (WHO) والأمم المتحدة (UN) إلى اتخاذ إجراءات عاجلة لمكافحة التفرقة العمرية بسبب آثارها السلبية على الرفاهية والوفاة المبكرة والتكاليف الصحية المرتفعة[7]. وأشار التقرير إلى أنه قد تُخصص موارد الرعاية الصحية الشحيحة أحياناً بناءً على العمر، مما يعني أن عمر الفرد قد يؤثر في الرعاية الصحية الأساسية.

3.2 التحيزات المرتبطة بتجهيز البيانات

3.2.1 تحيز اختيار العينة

يمكن أن ينتج التحيز في اختيار العينة عن طريق اختيار الأفراد أو المجموعات أو البيانات بطريقة لا تمثل فيها عينات الدراسة المراد تحليلها بشكل صحيح وعادل نتيجة ظهور أحد أنواع التحيزات التالية:

3.2.1.1 تحيز أخذ العينة

يحدث التحيز في أخذ العينات (Data Sampling) عندما تجمع البيانات بطريقة تكون فيها العينات المأخوذة من مجتمع ما أكثر من اللازم، بينما تكون العينات المأخوذة من مجتمع آخر أقل من المطلوب. وقد يكون هذا مقصوداً أو غير مقصود. ويعد هذا النوع من التحيز أحد أكثر أنواع تحيزات مجموعة البيانات شيوعاً. فعلى سبيل المثال، يمكن تغذية خوارزمية التعرف على الوجوه بمزيد من الصور للوجوه ذات البشرة الفاتحة مقارنة بالوجوه ذات البشرة الداكنة، مما يؤدي إلى ضعف الأداء في التعرف على الوجوه ذات البشرة الداكنة[8]. وبالتالي، يمكن أن يؤدي التحيز في أخذ العينات إلى زيادة نسبة الخطأ عند تعميم الخوارزميات المدربة.

3.2.1.2 تحيز المشاركة الطوعية

يشير تحيز المشاركة الطوعية إلى استجابة الناس طوعياً للمشاركة في مسح معين، مما قد يؤدي إلى انحراف في جمع البيانات، ولا يُتحصل في الواقع على نتائج تمثِّل جميع السكان. فالعديد من الدراسات الاجتماعية والاقتصادية والطبية وغيرها من المجالات تهدف للحصول على معلومات عن جميع السكان، عن طريق أخذ عينة عشوائية من السكان. والهدف من ذلك هو أن تمثل العينة المجتمع المعني بدقة، مع التأكيد على يتم الاختيار بطريقة تجعل العينة المختارة تمثل جميع السكان. وعندما تتكون العينة من متطوعين، فإن الخطر يكمن في أنهم لا يمثلون عامة السكان. ومن الأمثلة على تحيز المشاركة الطوعية ويسمى كذلك تحيز الاستجابة الطوعية هو عندما تنطبق الدراسة على الأشخاص من جميع مستويات الدخل ولكن هناك فقط متطوعين مشاركين من الفئة ذات الدخل المرتفع، هنا ستكون نتائج الدراسة مضللة.

3.2.1.3 تحيز التغطية

يحدث تحيز التغطية عندما تكون عينة البحث ليست ممثلة بشكل جيد للمجموعة المستهدفة بالدراسة؛ وقد يكون ذلك بسبب عدم إمكانية مشاركة فئة معينة في الدراسة. وفي معظم الحالات، ينتج تحيز التغطية عن الطريقة المستخدمة في أخذ العينة. ومن الأمثلة على هذا النوع من التحيز، هو استخدام الاستطلاع عبر الإنترنت. ولكن لا يزال هناك أشخاص لا يمكنهم الوصول إلى الإنترنت، أو لا يرغبون في استخدامه، ويمكن أن يؤدي ذلك إلى نتائج متحيزة لأن العينة قد لا تشمل بعض الفئات المهمة.

3.2.1.4 تحيز الاستبعاد

ينشأ تحيز الاستبعاد (Exclusion Bias) من البيانات التي تمت إزالتها بصورة غير ملائمة من مجموعة البيانات. حيث قد تستبعد بيانات معينة عن غير قصد، مما يؤدي إلى مجموعة بيانات متحيزة. فعلى سبيل المثال، إذا كان هناك شركة ما تستخدم 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 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 قد يكونون مؤهلين بنفس الدرجة ولكن لديهم خلفيات تعليمية أو عمل مختلفة غير ممثلة في مجموعة البيانات. نتيجة لذلك، قد تفوت الشركة توظيف مواهب متنوعة، مما قد يؤثر على إنتاجية الشركة.

3.2.2 تحيز القياس

تحيز القياس (Measurement Bias) يحدث عندما تتم المبالغة أو التقليل بشكل منهجي من القيمة الحقيقية للقياس. وكثيراً ما تشكل أخطاء القياس مصدراً كبيراً للتحيز في بيانات المسح[9]. فعلى سبيل المثال، عندما يتم أخذ عينات من أوزان المرضى، بينما الميزان المستخدم لذلك معيب، فإن البيانات التي سيتم تدوينها - دون أن تتم ملاحظة هذا الخلل - ستؤدي إلى أن تقوم الخوارزمية بتصنيف العملاء حسب أوزانهم إلى فئات خاطئة.

تحيز القياس (Measurement Bias) يحدث نتيجة عدم قياس أو تسجيل البيانات التي تم اختيارها بدقة. فعلى سبيل المثال، عندما تقوم شركة تأمين بأخذ عينات من أوزان العملاء المؤمنين لديها، بينما الميزان المستخدم لذلك معيب، فإن البيانات التي سيتم تدوينها - دون أن تتم ملاحظة هذا الخلل - ستؤدي إلى أن تقوم الخوارزمية بتصنيف العملاء حسب أوزانهم إلى فئات خاطئة[9]. وهناك ثلاثة أنواع من تحيز القياس وهي: تحيز التقاط البيانات، وتحيز الجهاز، وتحيز الوكيل.

3.2.2.1 تحيز التقاط البيانات

يحدث تحيز التقاط البيانات (Capture Bias) نتيجةً لبعض عادات الأشخاص في طريقة التقاط البيانات. فعلى سبيل المثال، عند إنشاء مجموعات بيانات تحتوي على صور وفيديو، فإن الصور أو مقاطع الفيديو قد تعكس التقنيات والعادات المستخدمة من قبل المصورين، وقد يميل بعض المصورين إلى التقاط صور للأشياء بطرق متماثلة؛ نتيجة لذلك قد تحتوي مجموعة البيانات على عرض لشيء معين من زوايا معينة فقط. وهذه الزوايا قد لا تساعد أن يكون الشيء الملتقط واضحاً بالشكل المطلوب الذي يساعد الخوارزمية على التعرف عليه بشكل جيد، بينما الأشياء الأخرى قد تكون ملتقطة بصورة واضحة تسهل من مهمة الخوارزمية في التعرف عليها.

3.2.2.2 تحيز الجهاز

تحيز الجهاز (Device Bias) هو مصدر آخر لتحيز القياس نتيجة للجهاز المستخدم لالتقاط مجموعات البيانات. فعلى سبيل المثال، قد تكون الكاميرات المستخدمة في التقاط الصور معيبة، مما يؤدي إلى ضعف جودة الصور و يصعّب على الخوارزمية التعرف على هذه الصور، وهذا بدوره يساهم في ظهور نتائج متحيزة.

3.2.2.3 تحيز الوكيل

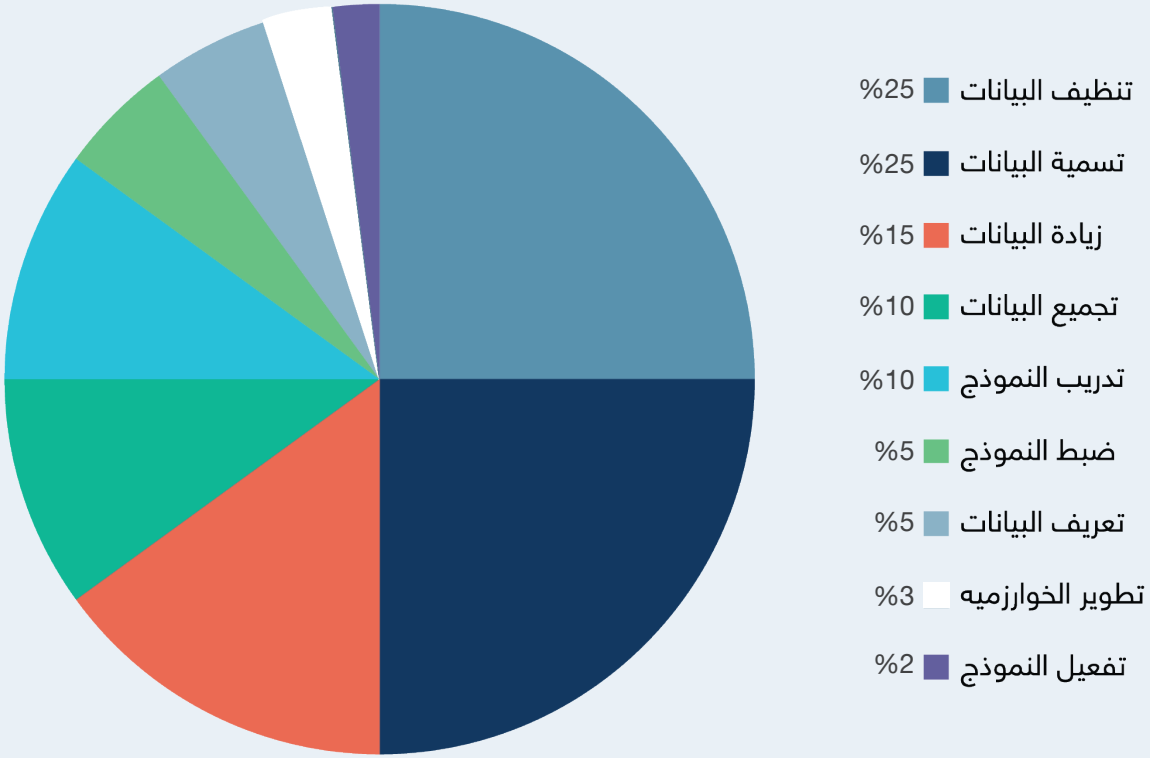
تحيز الوكيل (Proxy Bias) يحدث عند استخدام مؤشرات بديلة عوضاً عن القيم الحقيقية في إنشاء مجموعة البيانات مما يوجد نوعاً من التضليل. فعلى سبيل المثال، استخدام معدلات الاعتقال كمؤشر لمعدلات الجريمة، بينما يمكن أن يكون المعتقل بريئاً، أو استخدام عدد زيارات الطبيب كمؤشرات للحالات الطبية لانتشار مرض معين بينما بعض الزيارات التي تمت هي زيارات روتينية.

3.2.3 تحيز التسمية

تُعد تسمية البيانات (Label Bias) جزءاً من مرحلة ما قبل المعالجة في دورة أنظمة الذكاء الاصطناعي. إذ يتطلب تعريف البيانات الخام -مثل الصور والملفات النصية ومقاطع الفيديو- إضافة تسمية واحدة أو أكثر إلى تلك البيانات لتسهيل تعريفها للنماذج. وتُستخدم البيانات المسماة في تعلم الآلة الموجه، بينما تُستخدم البيانات غير المسماة في تعلم الآلة غير الموجه.

وتُشير شركة (Cognilytica) إلى أن 80% من الوقت المخصص لمشروع الذكاء الاصطناعي يتم شغله في مهام لها علاقة بالبيانات وإعدادها، 25% من هذا الوقت في تسمية البيانات [10]، انظر **شكل (3)** لمزيد من التفاصيل.

الشكل (3): نسبة الوقت المستغرق في تسمية البيانات يدوياً مقارنة بمهام مشروع الذكاء الاصطناعي الأخرى[10].



التدخل البشري في هذه المرحلة يزيد من احتمالية وجود الأخطاء. ففي ورقة بحثية، قام فريق بقيادة باحثين من مختبر علوم الحاسب والذكاء الاصطناعي (CSAIL) التابع لمعهد ماساتشوستس للتقنية بفحص 10 مجموعات بيانات رئيسية تشمل مجموعة البيانات (ImageNet)، ووجد بمجموعة البيانات هذه معدلات خطأ في عملية التسمية تصل إلى 6%[11].

3.2.3.1 طرق تسمية البيانات

تحيز التسمية يعتمد إلى حد كبير على طريقة تسمية البيانات وهو يعتبر خطوة حاسمة في تطوير نموذج تعلم الآلة عالي الأداء. لذا، من المهم أن تقوم الشركات بالنظر في طرق متعددة لتحديد أفضل نهج لتسمية البيانات. ولكون كل طريقة من طرق تسمية البيانات لها مزاياها وعيوبها، فإنه من المناسب التطرق لبعض الطرق لتسمية البيانات على النحو التالي:

- **تسمية البيانات داخلياً (Internal Labeling):** يؤدي استخدام خبراء علوم البيانات الداخليين إلى تبسيط عملية التتبع وتوفير دقة أكبر وزيادة الجودة. ومع ذلك، يتطلب هذا النهج عادةً مزيداً من الوقت وهو مناسب للشركات الكبيرة ذات الموارد الواسعة.

- **تسمية البيانات اصطناعياً (Synthetic Labeling):** البيانات الاصطناعية هي بيانات يتم إنشاؤها عن طريق الذكاء الاصطناعي على عكس البيانات الحقيقية التي تُجمع من العالم الواقعي، حيث غالباً ما تكون عملية جمع البيانات الحقيقية مكلفة أو متحيزة أو غير متوفرة أو غير قابلة للاستخدام بسبب قيود ولوائح الخصوصية. تقدر شركة جارتنر أنه بحلول عام 2030 ستتفوق البيانات الاصطناعية على البيانات الحقيقية في نماذج الذكاء الاصطناعي وستكون هي الأكثر استخداماً[12]، ومع ذلك، فإن تسمية البيانات اصطناعياً يتطلب قوة حوسبة كبيرة، مما قد يزيد التكلفة.

- **تسمية البيانات آلياً (Programmatic Labeling):** هي عملية استخدام البرامج النصية لتسمية البيانات آلياً لتقليل استهلاك الوقت والحاجة إلى التعليقات التوضيحية البشرية. ومع ذلك، فإن احتمال حدوث مشاكل فنية يتطلب أن يتم دمج تسمية البيانات الآلي مع فريق ضمان الجودة المختص، ليقوم هذا الفريق بمراجعة مجموعة البيانات أثناء عملية التسمية.

- **تسمية البيانات خارجياً (Outsourcing):** تكليف شركة خارجية متخصصة وذات خبرة في تسمية البيانات عوضاً عن إنجازها داخلياً، وقد تكون تكلفة الاستعانة الخارجية مرتفعة، خاصةً للمشاريع الكبيرة أو المُعقدة.

- **تسمية البيانات استعانةً بمصادر الحشود (Crowdsourcing):** يتم في هذه الطريقة توزيع مهام تسمية البيانات على مجموعة كبيرة من الأفراد عبر الإنترنت، غالباً من خلال منصات مخصصة. هذا النهج أسرع وأكثر فعالية من حيث التكلفة نظراً لقدرته على أداء المهام الصغيرة والتوزيع المستند إلى الويب. ومع ذلك، تختلف جودة الموظفين وضمان الجودة وإدارة المشاريع عبر منصات مصادر الحشود.

3.2.3.2 تحيز عدم التناسق

يحدث تحيز عدم التناسق (Inconsistencies Bias) عند عدم التناسق في عملية وضع التسمية. حيث لابد من وجود مصطلحات وتسميات موحدة للأشياء عند التسمية. ومن أمثلة تحيز عدم التناسق هو قيام المسمون المختلفون بتعيين تسميات مختلفة لنفس الكائن على سبيل المثال، العشب مقابل الثيل، لوحة فنية مقابل الصورة، وهكذا. مما قد يؤدي إلى عدم تعرف أنظمة الذكاء الاصطناعي على هذا الكائن.

3.2.3.3 تحيز المُسمِّي

تحيز المُسمّي (Annotator Bias) يحدث عندما تؤثر التحيزات الذاتية للمُسمّين على وضع التسميات. على سبيل المثال، إذا وضع أحد المُسمّين تسمية على صورة لسيارة على أنها "سيارة" والآخر على أنها نوع السيارة أو طرازها مثلاً (تويوتا)، فقد يؤدي ذلك إلى إرباك نظام الذكاء الاصطناعي. ومن أجل تلافي عدم التوافق في عملية التسمية يتم اللجوء إلى اتفاقية المسميين ((Inter-Annotator Agreement (IAA) وهي طريقة تُستخدم لتقييم جودة وضع العلامات من خلال قياس مستوى التوافق بين المصادر. ووفقاً لتقرير حديث لشركة (Accenture)، فإنه يلزم إجراء مزيد من المراجعة لضمان عدم انحراف البيانات إذا اكتشف القياس ضعف الإجماع بين المسميين[13].

3.2.3.4 تحيز تأثير نهاية الذروة

تحيز تأثير نهاية الذروة (The Peak End Effect Bias) هو نوع من التحيز المعرفي المرتبط بالذاكرة، إذ يحكم الناس على التجربة بناءً على شعورهم في ذروتها (أي النقطة الأكثر شدة) وفي نهايتها، بدلاً من الاعتماد على المجموع الكلي أو المتوسط لكل لحظة في التجربة. على سبيل المثال، قد يعطي بعض المسميين أهمية أكبر للجزء الأخير من المحادثة (بدلاً من المحادثة بأكملها) في تعيين تصنيف.

3.2.4 تحيز البقاء

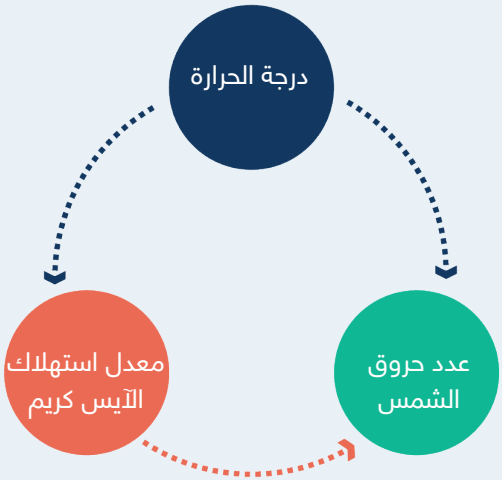
يحدث تحيز البقاء (Survivorship Bias) عندما يكون هناك ميل إلى تقييم النتائج الناجحة وتجاهل الإخفاقات. حيث يرسم هذا التحيز في العينة صورة أكثر وردية للواقع مما يتسبب في انحراف متوسط النتائج إلى الأعلى. فعلى سبيل المثال في القطاع المالي، قد يظهر تحيز البقاء عندما يكون هناك ميل لعدم تضمين الشركات الفاشلة في دراسات الأداء، لأنها اندثرت ولم يعد لها وجود. وهذا قد يتسبب في انحراف نتائج الدراسات لأنه تم فقط تضمين الشركات التي نجحت بما يكفي للبقاء على قيد الحياة.

3.2.5 تحيز الإرباك

يمكن أن ينشأ تحيز الإرباك (Confounding Bias) في نموذج الذكاء الاصطناعي إذا تعلمت الخوارزمية علاقات خاطئة بسبب عدم مراعاة جميع المتغيرات الموجودة في مجموعة البيانات، أو إذا فاتتها العلاقات ذات الصلة بين المعاملات والمخرجات المستهدفة. فعلى سبيل المثال، عندما تجمع بيانات لإيجاد العلاقة بين حروق الشمس واستهلاك الآيس كريم، سنجد أن ارتفاع استهلاك الآيس كريم يرتبط بارتفاع احتمالية الإصابة بحروق الشمس. لكن هل هذا يعني أن استهلاك الآيس كريم يسبب حروق الشمس؟.

هنا، المتغير المربك هو درجة الحرارة، فدرجات الحرارة المرتفعة تجعل الناس يأكلون المزيد من الآيس كريم ويقضون وقتاً أطول في الهواء الطلق تحت أشعة الشمس، مما قد يؤدي إلى مزيد من حروق الشمس. **شكل (4)** يبين العلاقة بين المتغير المربك والمتغيرات الأخرى.

الشكل (4): العلاقة بين المتغير المربك والمتغيرات الأخرى.



3.3 التحيزات المرتبطة ببناء النموذج

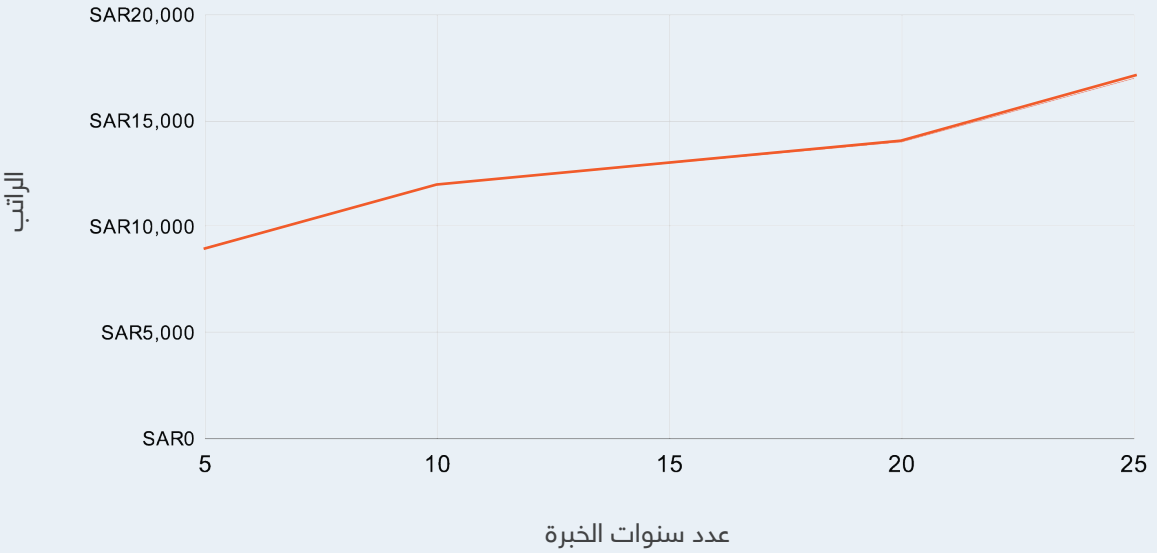
3.3.1 تحيز الخوارزمية

يمكن تعريف تحيز الخوارزمية (Algorithm Bias) على أنه التحيز الذي يحدث بسبب الخوارزمية بشكل متكرر دون أن يكون لمجموعة البيانات دورٌ. وهذا النوع من التحيز قد يكون نتيجة وجود قدرات محدودة للخوارزمية أو للنظام. على سبيل المثال البرمجيات التي تعتمد على العشوائية للتوزيع العادل للنتائج ليست عشوائية حقاً؛ فمن الممكن انحراف الخيارات نحو العناصر الموجودة في نهاية القائمة أو بدايتها، مما يؤدي إلى نتائج متحيزة.[14,15]

3.3.2 تحيز التجميع

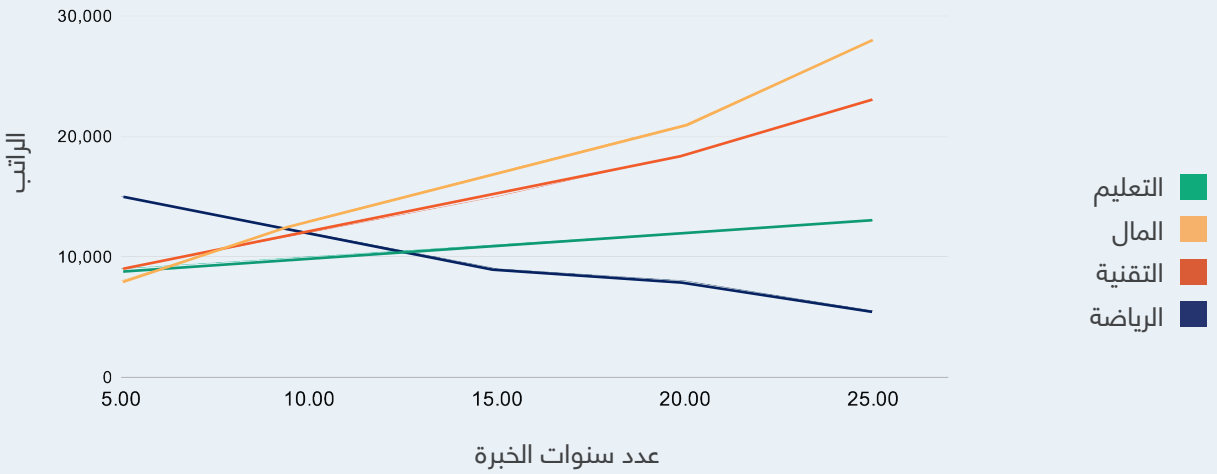
يحدث تحيز التجميع (Aggregation Bias) عند الافتراض الخاطئ بأن الاتجاهات التي تظهر في البيانات المجمعة للمتغيرات تنطبق أيضاً على بيانات المتغيرات الفردية، إذ يتم في بعض الأحيان تجميع البيانات لتبسيطها وتقديمها بطريقة سهلة، وهذا يمكن أن يؤدي إلى التحيز. ويوضح الرسم البياني في **شكل (5)** اتجاه رواتب الموظفين حسب عدد سنوات خدمتهم وذلك في عدة قطاعات: التعليم والمال والأعمال وتقنية المعلومات والرياضة.

الشكل (5): اتجاه رواتب الموظفين حسب عدد سنوات خدمتهم وذلك في قطاعات التعليم و المال والأعمال و تقنية المعلومات والرياضة[16].



من الواضح أن الراتب يزيد كلما زادت عدد سنوات الخدمة في الوظيفة. لكن عند إلقاء نظرة على البيانات التي تم تجميعها واستخدامها لإنشاء هذا المنحنى والموضحة في الرسم البياني **شكل (6)**، يُظهر الرسم البياني أن هذه المعلومة غير صحيحة بالنسبة للرياضيين، بل العكس هو الصحيح. حيث يوضح الرسم البياني أن الرياضيين قادرين على كسب رواتب عالية في وقت مبكر من حياتهم المهنية، بينما لا يزالون في أوج قوتهم الجسدية، ولكن دخلهم ينخفض مع مرور الوقت. ومما سبق يتضح أن تعميم النتيجة التي تم التوصل لها من تجميع البيانات من عدة قطاعات على كل قطاع بصورة منفصلة، قد يعطي نتائج مضللة، وهذا يجعل الخوارزميات متحيزة.

الشكل (6): الرياضيون قادرون على كسب رواتب عالية في وقت مبكر من حياتهم المهنية وهم لا يزالون في أوج قوتهم الجسدية، بينما ينخفض دخلهم مع مرور الوقت[16].



3.3.3 تحيز المتغير المحذوف

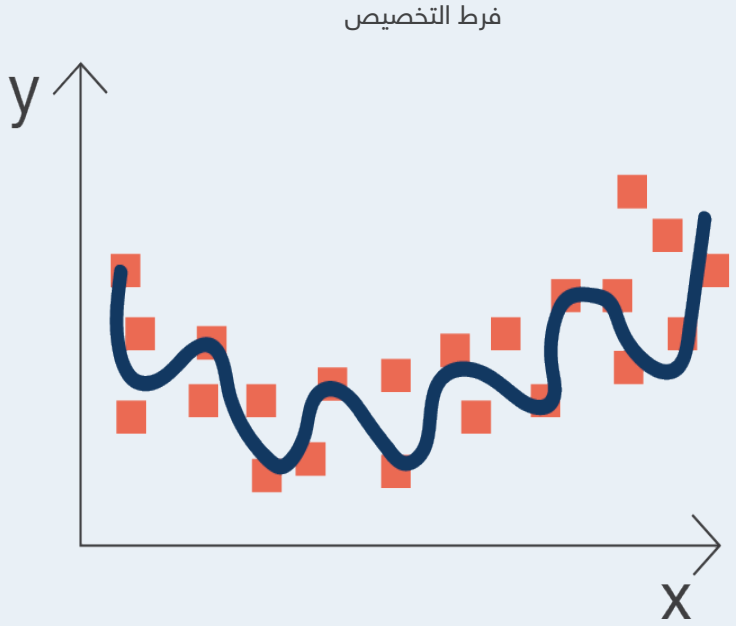
يحدث تحيز المتغير المحذوف (Omitted Variable) عند استبعاد معامل واحد أو أكثر من النموذج، ويحدث عادةً في نماذج تعلم الآلة التنبؤية. على سبيل المثال، مثال على تحيز المتغير المحذوف هو عندما يكون هناك دراسة تهدف إلى تحديد تأثير التعليم على الدخل ولكنها لا تتحكم في تأثير خبرة العمل. إذا كانت الخبرة العملية مرتبطة بشكل إيجابي بكل من التعليم والدخل، فإن إغفال هذا المتغير في التحليل سيؤدي إلى المبالغة في تقدير تأثير التعليم على الدخل. بعبارة أخرى، سيُظهر النموذج علاقة أقوى بين التعليم والدخل مما هو موجود بالفعل لأن تأثير تجربة العمل لا يتم أخذه بعين الاعتبار.

3.3.4 تحيز فرط التخصيص

فرط التخصيص (Overfitting) يحدث عندما يتطابق النموذج الإحصائي تماماً مع بيانات التدريب الخاصة به، مما يؤدي إلى عدم المحافظة على الدقة عند التعامل مع البيانات الجديدة، ويجعل النموذج غير قابل للتعميم على جميع البيانات.

إن معدلات الخطأ المنخفضة والتباين العالي هي مؤشرات لوجود فرط التخصيص. ومن أجل منع هذا النوع من السلوك، يتم عادةً وضع جزء من مجموعة بيانات التدريب جانباً من أجل عملية الاختبار للتحقق من عدم وجود "فرط التخصيص"[17]. وإذا تطابق النموذج بشكل وثيق جداً مع مجموعة التدريب، فإنه لا يمكنه أن يكون عام بحيث يتعامل مع البيانات الجديدة بشكل جيد. وهذا يؤدي إلى أن يكون النموذج غير قادر على أداء مهام التصنيف أو التنبؤ التي تم تصميمه من أجلها. الرسم البياني في **شكل (7)** يوضح هذه الفكرة.

شكل (7): محاولة تمرير النموذج بمعظم نقاط مجموعة بيانات التدريب بما في ذلك نقاط الضوضاء مما يسبب فرط التخصيص[18].

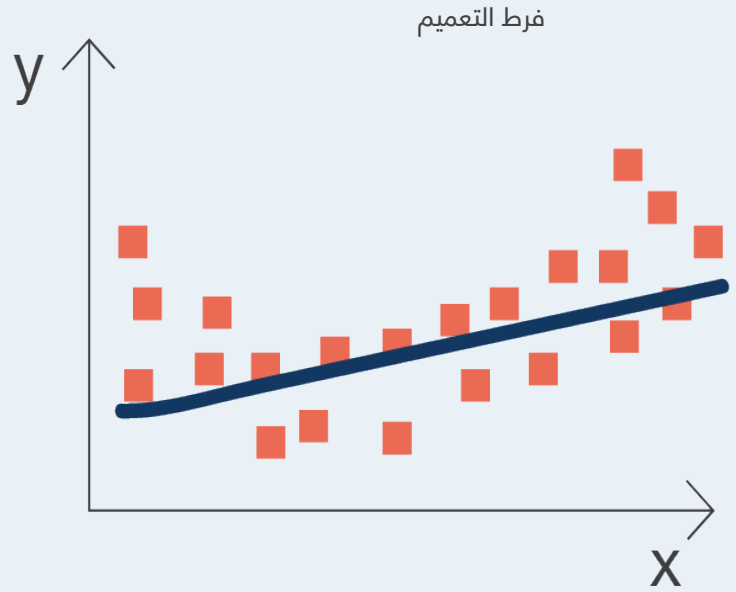


3.3.5 تحيز فرط التعميم

فرط التعميم (Underfitting) هو سيناريو معاكس لفرط التخصيص (Overfitting)، ويحدث عندما يتعذر على النموذج الإحصائي التطابق بشكل جيد مع بيانات التدريب الخاصة به، مما يؤدي إلى عدم المحافظة على الدقة عند التعامل مع البيانات الجديدة، وهذا يسبب ارتفاع معدل الخطأ في كل من مجموعة التدريب والبيانات التي لم ترى بعد (مرحلة تقييم النموذج)، بمعنى أن تكون دقة النموذج غير جيدة والتباين في مرحلة التحقق قليل.

الرسم البياني في **شكل (8)** يبين حالة فرط التعميم. إذ يبين أحد الأمثلة على حالة فرط التعميم التي يمكن أن تقع أثناء إنشاء النموذج، فمثلاً عندما محاولة مواءمة نموذج خطي مع بيانات الاتجاه السائد لها على شكل منحنى، حيث من الواضح أن هناك عدد من النقاط البعيدة عن هذا الخط، مما يتسبب في وجود نسبة خطاء عالية في مخرجات النموذج.

شكل (8): محاولة موائمة نموذج (ذو دالة خطية) بمجموعة بيانات تدريب ذات اتجاه سائد منحني مما يسبب فرط التعميم[18].



3.3.6 تحيز التمويل

يحدث تحيز التمويل (Funding Bias) عندما يؤثر مصدر التمويل لمشروع ما على نتائج النموذج أو تفسير هذه النتائج بطريقة تدعم الراعي المالي للمشروع. بحث أو دراسة معينة على النتيجة أو تفسير النتائج.

على سبيل المثال، إذا قامت شركة أدوية بتمويل تجربة سريرية لعقارها الخاص، فقد يكون هناك خطر من أن تصميم الدراسة أو تفسير النتائج يتأثر لصالح فعالية الدواء. في مثل هذه الحالات، قد تكون النتائج متحيزة لصالح مصدر التمويل بدلاً من البحث العلمي المحايد.

3.4 التحيزات المرتبطة باختبار النموذج

3.4.1 تحيز تسرب البيانات

يحدث تحيز تسرب البيانات (Data Leakage Bias) عندما يحدث تسرب للبيانات من القسم المخصص مثلاً للتقييم إلى القسم المخصص للتدريب، فالبيانات عادة تقسم إلى ثلاثة أقسام: قسم للتدريب وقسم مخصص للتقييم وقسم آخر مخصص للاختبار. وهذه الأقسام من المفترض ألا يكون بينها تداخل أو مشاركة حتى لا تؤثر على دقة التقييم والاختبار للنموذج حتى لا تعطي أرقام مضللة لدقة النموذج، وعادة ما تكون أعلى من الواقع في حال وجود تسرب للبيانات عند مشاركة المعلومات بين مجموعات بيانات التدريب والاختبار وهنا يحدث تحيز تسرب البيانات (Data Leakage Bias). ومن الأمثلة على ذلك، عندما يتم تدريب نموذج التعرف على الصور على البيانات التي تتضمن معلومات حول موقع الكائن داخل الصورة، فقد يبدو أن النموذج يعمل بشكل جيد عند اختباره على بيانات جديدة. ومع ذلك، قد تكون هذه الدقة الواضحة بسبب تضمين معلومات الموقع في بيانات التدريب، بدلاً من قدرة النموذج على التعرف بدقة على الكائن بناءً على ميزاته المرئية.

3.4.2 تحيزات التقييم البشري

يحدث تحيز التقييم البشري (Human Evaluation Bias) عندما يميل المقيِّمين البشريين إلى إدخال تحيزات في تقييماتهم للأفراد أو المجموعات بناءً على عوامل غير ذات صلة مثل العرق والجنس والعمر والحالة الاجتماعية والاقتصادية وما إلى ذلك. فعلى سبيل المثال عندما يقوم المقيمون بتقييم فرد أو مجموعة بشكل أكثر إيجابية بناءً على خاصية إيجابية لا علاقة لها بالمهمة الفعلية أو الأداء الذي يتم تقييمه. مثل أن يقوم المقيم بتقييم مرشح الوظيفة بشكل أفضل بسبب مظهره الجذاب أو سيرته الذاتية المثيرة للإعجاب، حتى لو لم تكن هذه العوامل ذات صلة بمتطلبات الوظيفة.

3.4.2.1 تحيز التأكيد

تحيز التأكيد (Confirmation Bias) هو نوع من التحيز المعرفي يحدث عندما يبحث الأفراد عن المعلومات أو يفسرونها أو يتذكرونها بطريقة تؤكد معتقداتهم أو فرضياتهم الموجودة مسبقًا، بينما يتجاهلون أو يستبعدون المعلومات التي تتعارض معهم. فعلى سبيل المثال في مجال الاستثمار قد يحضر الأشخاص بشكل انتقائي للمعلومات التي تؤكد معتقداتهم حول

استثمارات معينة. مثلاً، الشخص الذي يعتقد أن سهمًا معينًا هو استثمار جيد قد ينتبه فقط إلى القصص الإخبارية التي تؤكد هذا الاعتقاد ، بينما يتجاهل القصص الإخبارية السلبية التي تشير إلى أن السهم قد لا يكون استثماراً جيداً.

3.4.3 تحيز المراقب

يحدث تحيز المراقب (Observer Bias) عندما يقوم المختصون في بناء أنظمة الذكاء الاصطناعي بالتعديل في الأنظمة لتتناسب مع توقعاتهم أو رغبات المراقبين؛ وذلك بسبب عدم قبول أصحاب المصلحة بالنتائج التي لا تلبي توقعاتهم حتى وإن كانت صحيحة. على سبيل المثال، تم ملاحظة حدوث تحيز المراقب بشكل متكرر في دراسات ضغط الدم. حيث تم العثور على الأطباء الذين يقيسون ضغط الدم للمشاركين باستخدام مقاييس ضغط الدم الزئبقية لتقريب أو خفض القراءات إلى أقرب عدد صحيح. قد يحدث تحيز المراقب أيضًا إذا كان لدى الباحث فكرة مسبقة عما يجب أن يكون عليه ضغط الدم، مما يؤدي إلى تعديلات عشوائية في القراءات.

3.4.4 تحيز السبب والنتيجة

يحدث تحيز السبب والنتيجة (Cause-effect Bias) عندما يتم الخلط بين الارتباط والسببية. مثال شائع هو حول أكاديمي في الثمانينيات كان يبحث في معدلات الجريمة في مدينة نيويورك ووجد ارتباطاً قوياً بين كمية الآيس كريم التي يبيعها الباعة الجائلون ومعدلات الجريمة. حيث تم استنتاج أن تناول الآيس كريم يؤدي إلى زيادة معدلات الجريمة. لكن الاستنتاج المنطقي هو أن معدلات الجريمة كانت أعلى في الصيف، لكون الصيف هو الموسم الذي تكون فيه مبيعات الآيس كريم هي الأعلى.

3.5 التحيزات المرتبطة بنشر النموذج

3.5.1 تحيز انجراف البيانات

يحدث تحيز انجراف البيانات (Data Drift) بسبب التغير في توزيع بيانات الإدخال بمرور الوقت، و يقصد بتوزيع البيانات الكيفية التي يتم بها توزيع نقاط البيانات وترتيبها عبر مجموعة بيانات، ويتضمن معلومات حول نطاق قيم البيانات وتكرارها وأنماطها، وهذا يعني أن الخصائص الإحصائية لميزات الإدخال التي يستخدمها النموذج تتغير في مرحلة نشر النموذج عما كانت عليه في مرحلة التدريب، مما قد يؤدي إلى انخفاض في أداء النموذج. لذا لا بد من مراقبة النموذج وتقييمه بشكل دوري ليتم التأكد من مدى ملاءمته للبيانات الحالية، ومدى مواكبته للتغيرات.

3.5.2 تحيز حلقة التغذية الراجعة

يحدث تحيز حلقة التغذية الراجعة (Feedback Loop Bias) في مرحلة نشر نظام الذكاء الاصطناعي عندما تبدأ قرارات الذكاء الاصطناعي في التأثير على البيانات التي قد يتم تغذيتها به لاحقاً، مما يخلق دورة يمكنها تضخيم بعض التحيزات بمرور الوقت، حيث يمكن أن يؤدي هذا إلى تعزيز التحيزات الأولية لنظام الذكاء الاصطناعي، مما يجعلها أكثر وضوحاً وقد تؤدي إلى نتائج غير مرغوب فيها. فعلى سبيل المثال قد يتنبأ نظام الذكاء الاصطناعي المستخدم في الشرطة بمعدلات جريمة أعلى في مناطق معينة بناءاً على بيانات تاريخية متحيزة. مما قد يؤدي إلى زيادة وجود الشرطة في تلك المناطق، ومن ثم

المزيد من الحوادث المسجلة، مما يعزز اعتقاد النظام بأن هذه المناطق لديها معدلات جريمة أعلى، مما سيؤدي إلى ترسيخ هذا التحيز.

3.5.3 التحيز البيئي

يحدث التحيز البيئي (Environmental Bias) في مرحلة نشر نظام الذكاء الاصطناعي 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، فمثلاً قد تختلف الأجهزة التي يتم نشر نظام الذكاء الاصطناعي عليها عن بيئة التطوير، كالاختلاف في قوة المعالجة وجودة المستشعرات والاتصال، مما قد يؤثر على الأداء.

4. طرق التقليل من التحيزات

تشير معظم الأبحاث المرتبطة بطرق التقليل من التحيزات إلى صعوبة معالجة التحيزات كلياً، لكن من الممكن تقليلها بشكل كبير. والسعي لتقليل التحيزات في أنظمة الذكاء الاصطناعي هو خطوة في طريق تطوير أنظمة تتسم بالإنصاف وتتجنب التحيز، وبالتالي فهي محاولات جادة لدمج المبادئ الأخلاقية في تطبيقات الذكاء الاصطناعي وعملياته، وجزء رئيس في مفهوم "أخلاقيات الذكاء الاصطناعي".

ومن المهم ملاحظة أنه حتى لو تم اتباع أفضل الممارسات في تصميم الأنظمة وبناء النماذج، فإنها لن تكون كافية لإزالة مخاطر التحيز غير المرغوب فيه إذا كانت البيانات متحيزة. ومن هنا تأتي أهمية التوعية بتحيز البيانات حتى يمكن النظر في الأساليب التي تساعد على الحد من التحيز في تطبيقات الذكاء الاصطناعي وتعلم الآلة. كما تجدر الإشارة إلى أن كثيراً من الأدوات المتاحة حالياً للكشف عن التحيز ما زالت ناشئة ويمكن أن تكون مفيدة في حالات محددة فقط.

وفي دراسة حديثة نشرها المعهد الوطني للمعايير والتقنية في الولايات المتحدة (NIST)[19] تصف التحديات التي تواجهها قطاعات كثيرة وذلك عند الاستفادة من العامل البشري للمساعدة في تحسين قرارات أنظمة الذكاء الاصطناعي، وأوضح مثال على ذلك تقييم أحقية الحصول على قروض مالية في القطاع البنكي. وتهدف الدراسة إلى إنتاج تصميم مرجعي وأدلة إرشادية خاصة لكل قطاع وذلك لوصف منهجية التخفيف من التحيزات التي يمكن للمختصين الاستفادة منها عند نشر تطبيقات الذكاء الاصطناعي التي تقوم بأتمتة اتخاذ القرارات في قطاعاتهم. ويشتمل التصميم المرجعي على أربع مراحل للتخفيف من التحيزات وهي:

1. معالجة البيانات قبل استخدامها في النموذج للتأكد من عدم وجود تمييز أو تحيز.
2. التعديلات على الخوارزميات أثناء تدريب النموذج عند اكتشاف تمييز أو تحيز.
3. استخدام بيانات لم يتم تدريب النموذج عليها للتأكد من صحة مخرجات النموذج.
4. عرض النموذج الذي تم تطويره مع جميع التعديلات التي تم إدخالها على البيانات على عنصر بشري للتأكد من صحة جميع العمليات التي تم إجراؤها في الخطوات السابقة.

4.1 إرشادات عامة

الممارسات التالية مهمة لتكوين ثقافة في بيئة العمل تهيئ الموظفين لاكتشاف التحيزات، والإبلاغ عنها، والتعامل معها:

- **تنوع القوى العاملة:** من الضروري أن تكون القوى العاملة من خلفيات سكانية وثقافية واجتماعية مختلفة، مع تنوع في: العمر، والجنس، والتعليم، والخبرات، والسمات الشخصية، كون هذا التنوع يتيح للمنظمة أن تكون أكثر إبداعاً وقدرة على المنافسة، وتأتي هذه القوى العاملة المتنوعة بأفكار وتصورات مختلفة تساعد على اكتشاف الثغرات والتحيزات في الأفكار والتطبيقات.

- **توعية الموظفين:** توعية الموظفين بآثار التحيزات على المتضررين منها، وإجراء جلسات وحلقات نقاش مفتوحة حول المخاطر المحتملة للتحيز في دورة حياة بناء أنظمة الذكاء الاصطناعي. ومن المهم أن تكون التوعية بصفة دائمة وعلى فترات

تسمح بتغطية التطورات التقنية وإعطاء الموظفين فهماً أفضل عن التحيزات، وأن تشمل التوعية كبار المسؤولين في المنظمة، والمصممين والمطورين.

- **اختيار الموردين:** التأكد من أن أي مورد قد تم تصنيفه من الجهات التنظيمية المسؤولة عن البيانات والذكاء الاصطناعي وأن هذا المورد يطور ويدعم منتجات الذكاء الاصطناعي المسؤول ويقدم التدريب المنتظم بشأن التحيز للموظفين، وأن موظفيه حاصلون على شهادات مهنية احترافية في هذا الشأن.

- **البحث عن مصادر التحيزات والإبلاغ عنها:** حث الموظفين على البحث عن أنواع التحيزات في جميع دورة حياة بناء أنظمة الذكاء الاصطناعي، وتشجيعهم على عدم التردد في التعبير عن مخاوفهم عند ملاحظتهم لما يمكن أن يشكل تحيزاً يؤثر على القرارات التي تتخذها تلك الأنظمة.

- **وضع إجراءات للتخفيف من التحيزات:** يجب على المنظمات اعتماد إجراءات للتخفيف من التحيزات وذلك عند ملاحظتها من قبل الموظفين، أو من قبل الأدوات التي تساعد على اكتشافها، واتخاذ خطوات فورية لتخفيف تأثيرها على نظام الذكاء الاصطناعي.

- **حماية وحوكمة البيانات:** اعتماد إجراءات لحماية بيانات المستفيدين من أنظمة الذكاء الاصطناعي، والالتزام بالأنظمة والتشريعات التي تنظم جمع البيانات الشخصية ومعالجتها ومشاركتها بما يضمن المحافظة على خصوصية أصحاب هذه البيانات وحماية حقوقهم.

- **التدقيق الرسمي والمنتظم للخوارزميات:** يعد هذا الإجراء من أفضل الممارسات للكشف عن التحيزات وتخفيفها، وذلك عن طريق فرق عمل داخل المنظمة تكون مستقلة عن فريق التطوير، وتستطيع رفع التقارير بكل شفافية ووضوح عما وجدته أثناء إجراءات التدقيق. كما يمكن الاستعانة بشركات استشارية تقوم بعمليات التدقيق للقيام بهذه المهمة.

- **متابعة التطورات في هذا المجال:** تصدر العديد من المنظمات الدولية إرشادات خاصة بأنظمة الذكاء الاصطناعي المسؤول، كما تصدر العديد من الجهات الأكاديمية والشركات الاستشارية والتجارية البحوث والتقارير المرتبطة بهذا الموضوع، ويعد الاطلاع على هذه الإرشادات والبحوث والتقارير من الأمور المهمة للوقوف على آخر التطورات في طرق اكتشاف التحيزات والتقليل منها.

4.2 إرشادات متعلقة بدورة حياة بناء أنظمة الذكاء الاصطناعي

يستعرض هذا القسم مجموعة من الإرشادات المتعلقة بكل مرحلة من مراحل دورة حياة بناء أنظمة الذكاء الاصطناعي، وكيفية التعامل معها:

4.2.1 مرحلة تعريف المشكلة

- الحرص على الوصف الدقيق لمجموعة بيانات المستهدفين من الدراسة مثل: العمر، والجنس، والخلفية الثقافية، والعِرق، واللغة أو اللهجة، والمستوى التعليمي.

◁ وضع معايير محددة لتسمية البيانات يلتزم بها القائمون على هذه المهمة، وذلك لضمان الحصول على تسميات متسقة من المشاركين في هذه العملية.

4.2.2 مرحلة تجهيز البيانات

◁ تحديد الافتراضات الإحصائية الأساسية، مثل حجم العينة المطلوبة من المجتمع الإحصائي. وإذا تبين بعد خطوة جمع البيانات أن العينة صغيرة جداً، فمن المستحسن زيادة حجم العينة، حيثما أمكن ذلك.

◁ التأكد من أن حجم عينة البيانات تحتوي على القوة اللازمة لإعطاء استنتاجات معقولة باستخدام تحليل القوة (Power Analysis).

◁ توثيق جميع خطوات تصفية البيانات أو تعديلها، للمساعدة في تحديد مصادر التحيز مثل استبعاد خصائص معينة وهو ما يسمى بتحيز الاستبعاد (Exclusion Bias)، أو التسمية غير الصحيحة للبيانات والتي تعرف باسم تحيز التسمية (Label Bias).

◁ تقليل تحيز الاستبعاد (Exclusion Bias) من خلال البحث في كل خاصية قبل استبعادها. ويمكن القيام بذلك من خلال مساعدة الخبراء المختصين (SME's) الذين يمكنهم تحديد الخصائص الزائدة عن الحاجة أو باستخدام إحدى طرق تعلم الآلة مثل الغابة العشوائية (Random Forest) التي تنتج قائمة بأهمية المزايا.

◁ تقليل تحيز القياس (Measurement Bias) عن طريق تحديد القيم الشاذة (Outliers) - وهي قيم تختلف بشكل كبير عن متوسط البيانات الأخرى - ومن ثم حساب درجة تأثيرها على تغير النتائج باستخدام طرق إحصائية عديدة مثل مسافة كوك (Cook's Distance).

◁ موازنة مجموعات البيانات من خلال النظر في أساليب تقليل العينات (Down-Sampling) أو زيادة العينات (Over-Sampling)، باستخدام حزم (Python) تقوم بذلك تلقائياً مثل حزمة (SMOTE) لزيادة العينات، أو حزمة (Imblearn) التي تقوم بالأمرين معاً.

◁ حساب الارتباطات الزوجية (Pairwise Correlations) بين جميع المتغيرات التي تم تضمينها في النموذج، وهي طريقة إحصائية تُستخدم لتحديد العلاقة الخطية المتعددة مما يساعد في تحديد المتغيرات المربكة (Confounding Variables) التي تؤدي إلى التحيز المربك (Confounding Bias)، واتخاذ القرار بتضمين تلك المتغيرات في النموذج أو الاستغناء عنها.

4.2.3 مرحلة بناء النموذج

◁ اختيار الخصائص (Features) مهم لبناء نماذج جيدة، ومن الضروري تصفية الخصائص عديمة الصلة أو الزائدة عن الحاجة من مجموعة البيانات، مما يساعد في تقليل حدوث تحيز المتغير المهمل (Omitted Variable Bias)، وكذلك تحيز فرط التخصيص (Overfitting Bias).

◁ عندما تحتوي مجموعة البيانات على عدد كبير من الخصائص، يمكن استخدام تحليل المُكوِّن الرئيس (Principal Component Analysis) لتقليل عدد الخصائص في المكونات الرئيسية.

4.2.4 مرحلة اختبار النموذج

- أثناء اختبار النماذج للتأكد من صحتها، يمكن ملاحظة تحيز فرط التخصيص (Overfitting Bias)، وتحيز فرط التعميم (Underfitting Bias).
- تظهر مشكلة تحيز فرط التخصيص عندما يتعلم النموذج على بيانات التدريب بشكل مثالي وتكون نسبة الخطأ ضئيلة، ولكن أداءه ينخفض وتزداد نسبة الخطأ على بيانات الاختبار. وفيما يلي عدد من الأساليب التي يمكن استخدامها لتجنب فرط التخصيص:
- التوقف المبكر (Early Stopping): تسعى هذه الطريقة إلى إيقاف التدريب مؤقتاً قبل أن يبدأ النموذج في تعلم الضوضاء داخل النموذج. الهدف النهائي منها هو العثور على النقطة المثالية التي تحقق التوازن بين فرط التخصيص وفرط التعميم.
- التدريب على بيانات إضافية: يؤدي توسيع مجموعة التدريب لتشمل المزيد من البيانات إلى زيادة دقة النموذج من خلال توفير المزيد من الفرص لاستبعاد العلاقة السائدة بين متغيرات الإدخال والإخراج.
- زيادة البيانات (Data Augmentation): إضافة المزيد من مجموعات البيانات المشوشة باعتدال من مجموعات البيانات الأصلية مما يجعل النموذج أكثر استقراراً.
- الضبط (Regularization): إذا حدث فرط التخصيص بسبب تعقيد النموذج، فيمكن تقليل عدد الخصائص للحد من ذلك. وعملية الضبط تساعد في تحديد وتقليل الضوضاء داخل البيانات في حالة عدم معرفة أي المدخلات يجب حذفها أثناء عملية الاختيار. منهجيات الضبط مثل (L1 regularization, Lasso regularization, and dropout) يمكن الاستفادة منها لتحقيق ذلك.

إجراء اختبارات شاملة للنموذج في بيئة مشابهة للبيئة التي سيتم نشره فيها، وتقييم أداء النموذج باستخدام بيانات لم يسبق استخدامها في التدريب لضمان قدرته على التعامل مع بيانات جديدة.

4.2.5 مرحلة نشر النموذج

- تنفيذ نظام لمراقبة أداء النموذج بعد النشر، لضمان استمرارية الأداء العالي واكتشاف أي مشاكل بسرعة، وإجراء تقييمات دورية لأداء النموذج وتحديثه بناءاً على الملاحظات والبيانات الجديدة.
- تحديث النموذج بشكل دوري لضمان دقته وفعاليته لا سيما مع تغير الخصائص الإحصائية للبيانات وتغير البيئة التقنية والمتطلبات. كما ينبغي إنشاء خطة صيانة تتضمن جداول زمنية للتحديثات والاختبارات الدورية.
- توفير قنوات للتواصل ودعم المستخدمين والإجابة على استفساراتهم، وتشجيعهم على تقديم الملاحظات حول أداء النموذج ومخرجاته، وجمع هذه الملاحظات بشكل منتظم للاستفادة منها في تحسين النموذج باستمرار.
- وضع خطة طوارئ للاستجابة السريعة لأي علامات أو تقارير تشير إلى وجود تحيز أو مشكلة في النموذج.
- تقديم إرشادات واضحة للمستخدمين النهائيين حول كيفية استخدام النظام بشكل صحيح وآمن.

4.3 أدوات تقنية تساعد على اكتشاف التحيزات

(What-If Tool (WIT)) - شركة جوجل

◁ تسمح هذه الأداة التفاعلية المفتوحة المصدر للمستخدم بالتحقيق في نماذج تعلم الآلة للصور، مما يوفر فهماً لكيفية عمل النماذج في ظل السيناريوهات المختلفة وبناء تصورات غنية لشرح أداء النموذج. وتسمح خاصية اكتشاف التحيز للمستخدم بتحرير العينات يدوياً، ودراسة تأثير هذه التغييرات من خلال النموذج[20].

(FairML) - معهد ماساتشوستس للتقنية

◁ مجموعة أدوات شاملة لمراجعة النماذج التنبؤية من خلال تحديد الأهمية النسبية في النموذج التنبئي لتقييم الإنصاف (أو المدى التمييزي) لمثل هذا النموذج، مما يمكن المحللين من تدقيق النماذج التنبؤية المرهقة التي يصعب تفسيرها[21].

(AI Fairness 360) - شركة آي بي إم

◁ مجموعة أدوات مفتوحة المصدر قابلة للتطوير، تساعد في فحص التحيز في نماذج تعلم الآلة والإبلاغ عنها والتخفيف من حدتها طوال دورة حياة تطبيق الذكاء الاصطناعي[22].

(Fairlearn) - شركة مايكروسوفت

◁ مجموعة أدوات مفتوحة المصدر تُمكِّن علماء البيانات والمطورين من تقييم عدالة أنظمة الذكاء الاصطناعي وتحسينها. وتحتوي على لوحة معلومات تفاعلية وخوارزميات للتخفيف من الأضرار المتعلقة بالإنصاف قدر الإمكان[23].

(Aequitas) - جامعة شيكاغو

◁ مجموعة أدوات تدقيق التحيز مفتوحة المصدر، تقوم بمراجعة نماذج تعلم الآلة لاكتشاف التمييز والتحيز. تم تطويرها بواسطة مركز علوم البيانات والسياسة العامة بجامعة شيكاغو، ويمكن استخدامها لمراجعة تنبؤات أدوات تقييم المخاطر القائمة على تعلم الآلة (Predictions of Machine Learning based Risk Assessment Tools) لفهم أنواع مختلفة من التحيزات، واتخاذ قرارات مستنيرة بشأن تطوير ونشر مثل هذه الأنظمة[24].

accenture

(Teach and Test) - شركة أكسنتشر

◁ منهجية مصممة لمساعدة الشركات على بناء ومراقبة وقياس أنظمة الذكاء الاصطناعي الموثوقة داخل بنيتها التحتية الخاصة أو في السحابة. وتضمن المنهجية أن تنتج أنظمة الذكاء الاصطناعي القرارات الصحيحة، مع تجنب التحيزات، والمخاطر الأخلاقية، وعدم الامتثال للتعليمات التنظيمي[25].

Google

(ML-fairness-gym)- شركة جوجل

◁ مجموعة من الأدوات لبناء عمليات محاكاة تستكشف التأثيرات المحتملة على المدى الطويل لنشر أنظمة القرار القائمة على تعلم الآلة في البيئات الاجتماعية[26].

(Revise) - جامعة برنستون

◁ أداة تساعد في التحقيق في مجموعات البيانات المرئية، وتبرز التحيزات المحتملة. تساعد الأداة المستخدم من خلال اقتراح خطوات قابلة للتنفيذ يمكن اتخاذها للتخفيف من التحيزات التي تم الكشف عنها. الهدف الرئيسي هو معالجة مشكلة تحيز تعلم الآلة في وقت مبكر من عملية التدريب[27].

pwc

(Bias Analyzer) - شركة برايس ووترهاوس كوبرز

◁ خدمة تقدمها الشركة لعملائها تساعد على تحديد المخاطر المحتملة للتحيز ومراقبتها وإدارتها بشكل استباقي[28].

4.4 مبادرات دولية للتقليل من التحيزات

مسؤولية القطاع العام عن الخوارزميات - معهد (AI Now)

- إطار نموذجي موجه لمؤسسات القطاع العام لاستخدامه في تقييم الآثار الضارة المحتملة للخوارزميات[29].

أداة تقييم أثر الخوارزميات - كندا

- أداة لتقييم مخاطر أنظمة الذكاء الاصطناعي. والأداة عبارة عن استبانة من 48 سؤالاً حول المخاطر و33 سؤالاً للتخفيف منها. وتعتمد درجات التقييم على العديد من العوامل بما في ذلك أسلوب تصميم الأنظمة، والخوارزمية، ونوع القرارت التي تتخذها الأنظمة، والتأثير المحتمل من هذه القرارات، والبيانات التي تستخدمها هذه الأنظمة[30].

أداة التقييم الذاتي للأخلاقيات - بريطانيا

- أداة تقدم للباحثين إطار عمل سهل الاستخدام لمراجعة أخلاقيات مشاريعهم طوال دورة البحث. ويوفر التقييم الذاتي وسيلة مناسبة لتحديد القضايا الأخلاقية وتشكيل المناقشات المستقبلية. وتهدف العملية إلى دعم تقدير دقيق ومتسق لـ "المخاطر الأخلاقية" لمقترحات البحث[31].

أداة تقييم مخاطر الذكاء الاصطناعي - بريطانيا

- أداة تقوم بتقييم العديد من مخاطر الذكاء الاصطناعي بما في ذلك التحيز الذي يؤدي إلى التمييز، وعدم القدرة على التفسير، والهجمات الإلكترونية ذات التأثير الأكبر، والافتقار إلى الشفافية، وتلاشي الخصوصية في العدالة الجنائية، والخدمات المالية، والرعاية الصحية والاجتماعية، ووسائل الإعلام الرقمية والاجتماعية، والطاقة، والمرافق، ومن ثم تصنفها إلى ثلاثة مستويات: مخاطر عالية، ومتوسطة، ومنخفضة[32].

أداة التقييم الذاتي لمستوى أخلاقيات الذكاء الاصطناعي – دبي

- قائمة مرجعية لمساعدة المؤسسات المطورة للذكاء الاصطناعي أو المؤسسات المشغلة للذكاء الاصطناعي على التفكير في القضايا الأخلاقية المحتملة التي قد تنشأ خلال عملية التطوير، بما في ذلك مراحل الفكرة الأولية حتى صيانة النظام عندما يعمل بشكل كامل. المبادئ التوجيهية في أداة التقييم الذاتي اختيارية وليست إجبارية. تُستخدم الأداة لأغراض التقييم الذاتي فقط ولا يتم تدقيقها أو فحصها أو تنظيمها في الوقت الراهن[33].

5. التوصيات

- إعطاء مرحلة جمع البيانات واستعراضها مزيد من العناية، إذ يتضح من خلال الدراسة أن مرحلة جمع البيانات واستعراضها، هي أكثر مراحل أنظمة الذكاء الاصطناعي احتمالاً لظهور التحيزات.
- توخي الحذر والدقة عند جمع البيانات وبالتحديد عند أخذ العينات حيث أنها من المراحل بالغة الأهمية. والتحقق من أن بيانات التدريب التي يتم جمعها متوازنة، وممثلة للسكان قيد الدراسة تمثيلاً حقيقياً، كما يجب على المنظمات توثيق أساليبها في اختيار البيانات وتنقيتها.
- تعزيز الشفافية في عمليات تصميم وتطوير أنظمة الذكاء الاصطناعي، حيث لا بد أن يكون هناك وضوح في الأسس والمعايير التي يتم استخدامها في جميع مراحل بناء أنظمة الذكاء الاصطناعي.
- وضع معايير وقواعد موحدة ومناسبة لتسمية البيانات، لكي يتم تلافي عدم التوافق والاتساق في التسميات بين المسميين. ومن المناسب الاستعانة ببعض الأدوات الناضجة في هذا المجال مثل اتفاقية المسمين.
- تنويع فريق العمل المسؤول عن تطوير ونشر أنظمة الذكاء الاصطناعي وشموليته من حيث الجنس والعرق والثقافة والخلفيات العلمية ووجهات النظر والخبرات المختلفة. فرق العمل المتنوعة تسهم في الحد من التحيزات والتخفيف من حدتها من خلال جلب وجهات نظر وخبرات متنوعة.
- صيانة البنية التحتية لأنظمة الذكاء الاصطناعي لضمان الحفاظ على كفاءة هذه الأنظمة ووقايتها من تسرب التحيزات إليها، وخصوصاً تحيز القياس، لذا لا بد من عمل الصيانة الوقائية لأجهزة القياس والتأكد بشكل دوري من كونها تعمل بكفاءة.
- اعتماد إجراءات للتخفيف من تحيز البيانات عند ملاحظتها، وإجراءات لحماية بيانات المستفيدين، وعدم نشر التطبيقات إلا بعد مراجعتها من فرق الحوكمة المتخصصة والمدربة على البحث عن أنواع التحيزات في جميع دورات حياة بناء أنظمة الذكاء الاصطناعي، والإبلاغ عن أي مخاوف عند ملاحظتها. مع ضرورة وجود قنوات للإبلاغ عن التحيزات.
- مراقبة ومراجعة النماذج قيد التشغيل بشكل مستمر حيث أن عملية المراقبة والتدقيق والتحسين في أنظمة الذكاء الاصطناعي هي عملية مستمرة،
- الالتزام بالتدقيق الرسمي والمنتظم للخوارزميات للكشف عن التحيزات مبكراً وتخفيفها قبل استفحال ضررها.
- المتابعة المستمرة للبحوث المتسارعة والمتزايدة في هذا المجال.
- اختيار الموردين الملتزمين بتطوير تطبيقات "الذكاء الاصطناعي المسؤول" المعتمدين.
- دراسة التجارب الدولية المتعلقة بالذكاء الاصطناعي المسؤول ومتابعتها، والأدوات التي طورتها بعض الدول التي يمكن الاستفادة منها في التقليل من التحيزات.

6. المراجع

1. Fuller, J. B., Raman, M., Sage-Gavin, E. & Hines, K. Hidden Workers: Untapped Talent (2021). https://www.hbs.edu/managing-the-future-of-work/Documents/research/hiddenworkers09032021.pdf .

2. PricewaterhouseCoopers. Understanding algorithmic bias and how to build trust in AI **(2022).** https://www.pwc.com/us/en/tech-effect/ai-analytics/algorithmic-bias-and-trust-in-ai.html.

3. Schwartz, R. **et al. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence**. (2022) https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

4. Hao, K. This is how AI bias really happens—and why it's so hard to fix (2019). **MIT Technology Review** https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

5. Chander, R. S., Ajay. Biases in AI Systems (2021). https://cacm.acm.org/magazines/2021/8/254310-biases-in-ai-systems/fulltext.

6. Chu, C. H. **et al.** Digital Ageism: Challenges and Opportunities in Artificial Intelligence for Older Adults (2022). https://pubmed.ncbi.nlm.nih.gov/35048111/.

7. UN. Ageism is a global challenge (2021). https://www.who.int/news/item/18-03-2021-ageism-is-a-global-challenge-un.

8. Buolamwini, J. & Gebru, T. Gender shades Intersectional accuracy disparities in commercial gender classification. (2018). https://www.media.mit.edu/events/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/

9. Pablo Celhay, Bruce D. Meyer, Nikolas Mittag, What leads to measurement errors? (2024). https://www.sciencedirect.com/science/article/pii/S030440762300297X

10. **Cognilytica**. Data Preparation & Labeling for AI (2020). https://www.cognilytica.com/document/data-preparation-labeling-for-ai.

11. MIT CSAIL. Major ML datasets have tens of thousands of errors (2021). https://www.csail.mit.edu/news/major-ml-datasets-have-tens-thousands-errors.

12. Gartner. Is Synthetic Data the Future of AI? (2022). https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai.

13. Accenture. Whitehouse, M. How to Identify and Mitigate Bias in Federal AI (2020).

https://www.accenture.com/us-en/blogs/federal-viewpoints/how-to-identify-and-mitigate-bias-in-federal-ai

14. ACM. Srinivasan, R. & Chander, A. Biases in AI systems (2021).

15. GOV.UK. Rovatsos, M., Mittelstadt, B. & Koene, A. Landscape Summary: Bias in Algorithmic Decision-Making (2019). https://assets.publishing.service.gov.uk/media/5d31c30a40f0b64a8099e21d/Landscape_Summary_-_Bias_in_Algorithmic_Decision-Making.pdf

16. Statology. ZACH. What is Aggregation Bias? (Explanation & Example) (2020). https://www.statology.org/aggregation-bias/.

17. IBM. What is Overfitting? (2021). https://www.ibm.com/cloud/learn/overfitting.

18. Amazon. Model Fit: Underfitting vs. Overfitting - Amazon Machine Learning (2022). https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html.

19. NIST. A. Mitigating AI/ML Bias in Context (2022). https://csrc.nist.gov/pubs/pd/2022/11/09/mitigating-ai-ml-bias-in-context/final

20. Google. Using the What-If Tool | AI Platform Prediction (2022). https://cloud.google.com/ai-platform/prediction/docs/using-what-if-tool

21.MIT. Adebayo, J. A. ToolBox for diagnosing bias in predictive modeling (2016). https://dspace.mit.edu/handle/1721.1/108212

22.IBM Research. Introducing AI Fairness 360, A Step Towards Trusted AI (2018). https://www.ibm.com/blogs/research/2018/09/ai-fairness-360

23.Microsoft. Fairlearn: A toolkit for assessing and improving fairness in AI. (2020). https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf

24.Rayid. Aequitas. Data Science and Public Policy (2019). https://arxiv.org/pdf/1811.05577.pdf

25.Accenture. Teach and Test. https://ttp.accenture.com/ttp/TeachandTest.

26.Google. ML-fairness-gym: A Tool for Exploring Long-Term Impacts of Machine Learning Systems (2020). https://ai.googleblog.com/2020/02/ml-fairness-gym-tool-for-exploring-long.html

27.Wang, A.,Alexander, L., Zhang, R., Kleiman A., REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets (2022). https://par.nsf.gov/servlets/purl/10373920

28.PWC. Bias Analyzer (2022).

https://www.pwc.com/us/en/services/consulting/cloud-digital/data-analytics/artificial-intelligence/bias-analyzer.html

29.Ortolano, L. & Shepherd, A. ENVIRONMENTAL IMPACT ASSESSMENT: CHALLENGES AND OPPORTUNITIES (2012). https://www.tandfonline.com/doi/abs/10.1080/07349165.1995.9726076

30.Secretariat, T. B. of C. Algorithmic Impact Assessment Tool (2021). https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

31.**UK Statistics Authority**. Ethics Self-Assessment Tool (2022).

https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/

32. GOV.UK. AI Barometer (2021**).** https://www.gov.uk/government/publications/ai-barometer-2021.

33.Digital Dubai. AIEthics Self Assessment (2022). https://www.digitaldubai.ae/self-assessment

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

SDAIA.GOV.SA SDAIA_SA SDAIA.SAUDI SDAIA-KSA