

جامعة ميدأوشن
كلية تقنية المعلومات
مشروع خطة أطروحة للتسجيل لدرجة الدكتوراة

العنوان باللغة العربية:

تطوير آلية للتعامل مع الذكاء الاصطناعي في البيئات غير الخاضعة للرقابة نتيجة التحيز
من تقييم المخاطر إلى التخفيف والتعافي

العنوان باللغة الإنجليزية:

Developing a framework for Artificial Intelligence in uncontrolled environments due to bias, From Risk Assessment to Mitigation and Recovery

اعداد:

الباحث: علي حسن الشهري

المملكة العربية السعودية

الرقم الاكاديمي: 20223006415

رسالة مقدمة لجامعة ميدأوشن لاستكمال متطلبات نيل درجة الدكتوراة في تقنية المعلومات

By:

Researcher: Ali Hassan Alshehri

Kingdom of Saudi Arabia

Submitted to Midocean University in
partial fulfillment of the requirements for the degree of Doctor of Philosophy in
IT

قال تعالى في محكم التنزيل

وما أوتيتم من العلم إلا قليلا

صدق الله العظيم

اهداء

أهدي هذا الجهد البحثي للمهتمين بالذكاء الاصطناعي وأمنه ولوالدي رحمهما الله وآمل أني قد وفيت بما وعدت ولعائلي العزيزة لدعمهم وتحفيزهم المستمر.

الباحث

الشكر والتقدير

للمشرف على هذه الدراسة د.محمد الخولي على حرصه وارشاده وتوجيهاته

للأستاذة والاستاذات الدكاترة الأفاضل أعضاء اللجان البحثية والعلمية بجامعة ميدأوشن الذين قدموا النصح والمشورة لتطوير وتحسين مخرجات الدراسة

لجامعة ميدأوشن وفريق عملها الداعم

الملخص:

مع استمرار تطور تكنولوجيا الذكاء الاصطناعي، تتزايد المخاوف والقلق بشأن المواقف التي تعمل فيها هذه التقنيات خارج نطاق الإشراف البشري. تقدم هذه الدراسة نهجًا للتفاعل مع الذكاء الاصطناعي في مثل هذه السياقات الحرجة في ظل آليات التحكم. تهدف الدراسة إلى تحديد المخاطر، ووضع خطط لاحتوائها، ووضع خطط طوارئ للتهديدات التي تشكلها عمليات الذكاء الاصطناعي غير الخاضعة للرقابة بسبب التحيز. من خلال البحث في الجوانب النظرية المقارنة، والتنبؤ بالمخاطر، وبناء سيناريوهات للتهديدات المحتملة في سياق الخدمات التي يقدمها الذكاء الاصطناعي، يساعد هذا البحث في مناقشة سلامة الذكاء الاصطناعي من خلال تحليل أساليب تقييم المخاطر، والإجراءات التحضيرية، والجوانب الأخلاقية للتحكم في الذكاء الاصطناعي. تساعد هذه الدراسة في تطوير إطار تنظيمي عملي يهدف إلى تقليل التحيز في أنظمة الذكاء الاصطناعي من خلال إنشاء طريقة موثوقة للكشف عن التحيز. بالإضافة إلى ذلك، تقترح استراتيجيات للتخفيف من التحيز والتعافي منه، وتعزيز سلامة ونزاهة وموثوقية تقنيات الذكاء الاصطناعي، لا سيما فيما يتعلق بالحفاظ على توافقها مع القيم الإنسانية ومعايير الأمن.

الكلمات المفتاحية: إدارة مخاطر الذكاء الاصطناعي، أمن الذكاء الاصطناعي، التحكم في الذكاء الاصطناعي، تخطيط التعافي، طرق التحكم، أخلاقيات الذكاء الاصطناعي ولوائحه.

Abstract:

As AI technology continues to evolve, so too grows the fears, concerns are growing about situations in which these technologies operate outside of human supervision. This study presents an approach to interacting with AI in such critical contexts under control mechanisms. The study aims to identify risks, develop containment plans, and develop contingency plans for threats posed by uncontrolled AI operations due to bias. By research comparison theoretical aspects, risk prediction, and scenario building for potential threats in the context of AI-provided services, this research helps in the discussion of AI safety by analyzing risk assessment methods, preparatory procedures, and ethical aspects of AI control. This study helps develop a practical regulations framework aimed to minimizing bias in AI systems by establishing a reliable bias detection method. Additionally, it proposes strategies to mitigate and recovery from bias, enhancing the safety, fairness, and reliability of AI technologies, particularly with regard to maintaining their compatibility with human values and security standards.

Keywords: AI risk management, AI security, AI control, recovery planning, control methods, AI ethics and regulations.

المحتويات:

الفصل الأول: المقدمة والإطار النظري ومشكلة البحث

الفصل الثاني: الإطار النظري للدراسة (مراجعة الأدبيات ذات العلاقة)

الفصل الثالث: الطريقة والإجراءات (المنهجية)

الفصل الرابع: نتائج الدراسة

الفصل الخامس: خاتمة الدراسة

المستخلص والتوصيات

المراجع

الملاحق

الفصل الأول: المقدمة والإطار النظري

مقدمة الدراسة:

يزعم مطورو نظام الذكاء الاصطناعي العام "مانوس" (نظام الذكاء الاصطناعي العام)، الذي ظهر لأول مرة في ٦ مارس ٢٠٢٥، أنه أول عامل ذكاء اصطناعي مستقل تمامًا في العالم، لا يساعد البشر فحسب، بل يحل محلهم في وظائف معينة. بناءً على هذه المعلومات، يمكننا القول إن الذكاء الاصطناعي قد انتقل من أتمتة المهام الأساسية إلى أنظمة تعمل تلقائيًا (مما يثير مخاوف الناس من أن تخرج الأمور عن السيطرة). تتناول هذه الدراسة سلوك الذكاء الاصطناعي، وتركز على ثلاثة عناصر رئيسية: الكشف (لتحديد التهديدات المحتملة)؛ والتخفيف (الإجراء الذي يعد حل مناسب)؛ والتعافي (لإعادة الأمور إلى نصابها) بسبب التحيز الذي قد ينشأ داخل النظام. علاوة على ذلك، تم في هذه الدراسة التطرق إلى عدم موثوقية الذكاء الاصطناعي فقد نشر موقع IBM مقالة للمؤلفات (Alice Gomstyn، Amanda McGrath، Alexandra Jonker) عن دراسة قام بها الباحثون Mesut Cicek و Dogan Gursoy و LuLu من جامعة واشنطن الوطنية أظهرت الدراسة زيادة في حالات نقص الثقة حيث وجد استطلاع أجري في عام ٢٠٢٣ أن أكثر من ٤٠% من قادة الأعمال أشاروا إلى مخاوفهم بشأن مصداقية الذكاء الاصطناعي وفي الوقت نفسه، أظهر المستهلكون أيضًا عدم ثقتهم في الذكاء الاصطناعي: فقد وجدت دراسة أجريت عام ٢٠٢٤ أن تضييق مصطلح "الذكاء الاصطناعي" في ملصق المنتج يمكن أن يجعل المتسوقين أقل ميلًا لشراء هذا المنتج، وهذا النقص في الثقة يمكن أن يحدث مع أنظمة الذكاء الاصطناعي التي تعمل مثل التفكير البشري، والذي يمكنه معالجة إجراءات مثل اختراق الأنظمة أو رفض الوصول، بالإضافة إلى تقنيات التزييف العميق المتقدمة. وفي ظل هذه المعطيات يصبح الذكاء الاصطناعي في وضع يجمع بين الأهمية والمخاطر لذا علينا اتباع قواعد وإرشادات صارمة في هذا الصدد، ويجب ألا نغفل أبدًا عن الأساليب الأكثر فعالية التي تتحكم في الذكاء الاصطناعي.

يُثير النمو السريع للذكاء الاصطناعي، وخاصةً نحو الآلات التي تُنافس ذكاء الإنسان، صعوبات حقيقية. ماذا يحدث عندما تبدأ هذه الأنظمة بفعل أشياء لم تُخطط لها؟ أجرت هذه الدراسة بحثًا مُباشرًا، مستندة إلى عدد من السيناريوهات القائمة على التحيز سواء كان ادراكيًا أو تقنيًا حيث يُمكن تحليل مستوى التحيز من خلال التغذية الراجعة للحفاظ على الاستقرار والأمن، بالإضافة إلى تعزيز إعدادات السلامة التي قد تُمثل مكابح طوارئ لنا (يامبولسكي، ٢٠١٦؛ هادفيلد-مينيل وآخرون، ٢٠١٧). مع ذلك، هناك عقبة: فنحن نفتقر إلى الخطط التي تربط بين اكتشاف المشكلات مُبكراً، وإيقافها بأساليب التحكم، يتحدث بعض الباحثين، مثل سواريس وآخرون (٢٠١٥)، عن سهولة إيقاف الذكاء الاصطناعي. ويُشير بوستروم (٢٠١٤) إلى مخاطر كبيرة، مثل أن يُدمرنا الذكاء الاصطناعي دون استراتيجيات واضحة لإيقافه.

مشكلة الدراسة:

المشكلة البحثية الأساسية:

تشهد أنظمة الذكاء الاصطناعي تطوراً متسارعاً، وهو ما يُثير تحديات جوهرية تتعلق بإمكانية فقدان السيطرة على هذه الأنظمة نتيجة التحيز. وتتمثل الإشكالية المركزية في غياب إطار منهجي متكامل للكشف عن هذه التحيزات، وتخفيف آثارها، واستعادة السيطرة عند الضرورة. هذا النقص يعرض مصداقية أنظمة الذكاء الاصطناعي وأمنها للخطر، ويقوّض ثقة المجتمع في هذه التقنيات المتقدمة.

الأبعاد الجوهرية للمشكلة:

١. أزمة الموثوقية

تشير الأدبيات التجريبية الحديثة (Cicek et al., 2023) إلى تزايد التشكك المؤسسي في مصداقية أنظمة الذكاء الاصطناعي، حيث أعرب نحو ٤٠% من قادة الأعمال عن قلقهم حيال مدى موثوقية هذه الأنظمة. كما تكشف الدراسات السلوكية (٢٠٢٤) عن تراجع واضح في تقبل المستهلكين للمنتجات التي تعتمد على الذكاء الاصطناعي، مما يعكس أزمة ثقة أخذت في الاتساع.

٢. مخاطر تحيز الأنظمة:

يؤدي التحيز الكامن داخل أنظمة الذكاء الاصطناعي إلى مجموعة من التحديات، تشمل:

- تشوّه عمليات اتخاذ القرار، كما هو الحال في التمييز الخوارزمي كما تشير (Timedjehdine N, ٢٠٢٥)، مخاطر التحيز الخوارزمي في الذكاء الاصطناعي)
- سلوكيات غير متوقعة قد تتضمن اختراق الأنظمة أو إنتاج محتوى مزيف (مثل التزييف العميق). هندريكس د. وآخرون، ٢٠٢٣، نظرة عامة على مخاطر الذكاء الاصطناعي الكارثية)
- تهديدات وجودية محتملة كما ورد في تحليل (Bostrom (2014 حول مخاطر الذكاء الاصطناعي المتقدم.

٣. القيود في آليات التحكم

تعاني أنظمة الذكاء الاصطناعي الحالية من محدودية في آليات التحكم والمراقبة، ويتجلى ذلك في:

- غياب أنظمة إنذار مبكر قادرة على اكتشاف التحيز في مراحله الأولية.
- محدودية فعالية استراتيجيات الطوارئ، مثل آليات "مكابح الطوارئ" (Yampolskiy, 2016)
- قصور في بروتوكولات الاسترداد بعد التدخل مما يعيق استعادة السيطرة الفعالة على النظام عند حدوث خلل.

تناقش الدراسة الأسئلة التالية:

- ما هي المخاطر الدقيقة التي تجعل الذكاء الاصطناعي يخرج عن السيطرة نتيجة مشاكل قائمة على التحيز؟
- كيفية السيطرة على أنظمة الذكاء الاصطناعي واستدراك التحيز قبل حدوثه.
- كيف تسهم الأساليب المستخدمة في تحسين المستوى الأخلاقي للنظام؟
- بعد استخدام هذه الأساليب والتقنيات، كيف نعود إلى حياتنا الطبيعية ونحافظ على أخلاقيات العمل؟

هذه الأمور مهمة لأن الأمور الحياتية، مثل إدارة الذكاء الاصطناعي لتداول الأسهم أو الطائرات العسكرية بدون طيار أو وسائل النقل، حيث تشير الأدبيات إلى أن غياب أنظمة الإنذار المبكر قد تؤدي إلى انخيار أنظمة الذكاء الاصطناعي إذا لم يتم التدخل بشكل عاجل وفعال.

تهدف الدراسة إلى ربط الأفكار النظرية بإستراتيجيات عملية، من خلال أساليب وتقنيات لاستعادة السيطرة، مع تفعيل الحوكمة الأخلاقية للحفاظ على فعاليتها ونزاهتها.

أهمية الدراسة:

الأهمية النظرية:

تتعمق هذه الدراسة في سلامة الذكاء الاصطناعي مع التركيز على التحيز كمؤثر على السلامة، مع إعطاء أساليب المراقبة والتفاعل مع الذكاء الاصطناعي دورًا محوريًا في خطة ثلاثية الأجزاء: الكشف، والتخفيف، والتعافي، كما تفتح نقاشًا حول الحوكمة الأخلاقية للذكاء الاصطناعي.

الأهمية التطبيقية:

تقدم الدراسة أسلوب تحسين للتعامل مع خروج الذكاء الاصطناعي عن السيطرة بسبب التحيز وهو ما يضيف لآليات متابعة أنظمة الذكاء الاصطناعي ومراقبتها وإعادة بناء الثقة، والتأكد من عدم خروج هذه الأنظمة عن السيطرة.

أهداف الدراسة:

تعمل الدراسة على تطبيق إجراءات وأساليب وتقنيات للتحكم في الذكاء الاصطناعي الخارج عن السيطرة. وتستند إلى مبادئ تتقاطع فيها القضايا الأخلاقية والتقنية، وذلك لتغطية طبيعة هذا النوع من تطبيقات الذكاء الاصطناعي، وتسعى الدراسة لضمان بقاء النتائج تحت السيطرة. وتشمل أهداف الدراسة الرئيسية ما يلي:

- تحديد نقاط الفشل وتقييم المخاطر لاستخدام الذكاء الاصطناعي.
- دراسة المواقف المحتملة التي قد ينخرط فيها الذكاء الاصطناعي في سلوك غير مقصود أو ضار أو عدائي.
- البحث عن علامات تحذيرية تشير إلى هروب الذكاء الاصطناعي من السيطرة، مثل التغيير أو الخداع أو التهرب من القيود.
- تحديد الثغرات في أطر حوكمة الذكاء الاصطناعي الحالية وهيكل صنع القرار.
- تسليط الضوء على الحلول وفقاً لما هو متاح لكل نظام للذكاء الاصطناعي.
- تقييم موثوقية وكفاءة وأمان طرق الكشف والتخفيف والتعافي.
- يجب أن تكون آليات الكشف والتخفيف والتعافي لتشغيل الذكاء الاصطناعي متوافقة مع مبادئ الذكاء الاصطناعي الأخلاقية، بحيث تُسبب أقل قدر من الضرر للمستخدمين والبنية التحتية.
- البحث في مخاطر تفعيل الحلول على أنظمة الذكاء الاصطناعي، بما في ذلك الرعاية الصحية، والنقل، والتقنيات ذاتية التشغيل.
- تقييم أطر الحوكمة لتحديد متى ومن يجب عليه تفعيل الحلول.
- استخدام تدابير السلامة من الأعطال والتكرار ودور الرقابة البشرية وأنظمة المراقبة الآلية في استراتيجيات احتواء الذكاء الاصطناعي.
- وستضمن الدراسة الاستمرار في استخدام أسلوب SMART (محدد، قابل للقياس، قابل للتحقيق، ذو صلة، محدد زمنيًا).
- تحديد وتحليل المخاطر المحتملة نظريًا عند خروج الذكاء الاصطناعي عن السيطرة بسبب التحيز من خلال قائمة تحقق شاملة ومنظمة، يتم التحقق منها من خلال حالات واقعية، مرتبطة بالسلامة، خلال 4 أشهر.

- بناء طريقة للكشف المبكر عن المشكلات، باستخدام أدوات الكشف، خلال ٦ أشهر.
- دراسة أساليب إيقاف الذكاء الاصطناعي الخارج عن السيطرة، والمثبتة من خلال سيناريوهات حالات عملية، مع التركيز على معايير السلامة، خلال ٨ أشهر.
- تتبع التحسينات في الذكاء الاصطناعي، باستخدام إجراءات التحكم والإجراءات الأخلاقية لتطوير وتنفيذ استراتيجيات استرداد فعالة، خلال شهرين.

مفاهيم الدراسة:

الذكاء الاصطناعي: ويعرف بأنه التقنية التي تمكن الحواسيب والآلات من محاكاة التعلم البشري والفهم وحل المشكلات واتخاذ القرارات والإبداع والاستقلالية كما يوضح Kavlakoglu E, Stryker C

حوكمة الذكاء الاصطناعي: وهي مجموعة من الإجراءات والحواسخ تهدف لحماية و ضمان بقاء أدوات وأنظمة الذكاء الاصطناعي آمنة وأخلاقية (Kavlakoglu E, Stryker C)

التحيز في الذكاء الاصطناعي : هو خطأ يقع عندما يتم تغذية نماذج تدريب الذكاء الاصطناعي ببيانات مغلوطة تعكس أحكاماً مسبقة، أو يتم ترجيح بيانات على أخرى، ما يؤدي إلى نتائج منحرفة، وأخطاء تحليلية (معاد م 2020 كيف يحدث تحيز الذكاء الاصطناعي وما هي حلوله الممكنة؟)

معايير السلامة في الذكاء الاصطناعي: هي الممارسات والمبادئ التي تساعد على ضمان تصميم تقنيات الذكاء الاصطناعي واستخدامها بطريقة تفيد البشرية وتقلل من أي ضرر أو نتائج سلبية محتملة (Jonker A, McGrath A(2024) What is AI safety

وتركز هذه الدراسة على إشكالية فقدان السيطرة على أنظمة الذكاء الاصطناعي نتيجة التحيزات المدججة، سواء كانت إدراكية أو تقنية، وما يترتب على ذلك من آثار على الموثوقية، والسلامة، والحوكمة الأخلاقية لهذه الأنظمة.

تنطلق الدراسة من ثلاث ركائز مفاهيمية رئيسية:

أزمة الموثوقية:

يتمثل هذا البعد في التراجع الملحوظ في ثقة المؤسسات والمجتمع في أنظمة الذكاء الاصطناعي، نتيجة حوادث أو مؤشرات على قرارات خاطئة أو متحيزة، ما يهدد اعتماد هذه الأنظمة في المجالات الحساسة.

مخاطر تحيز الأنظمة:

يشتمل هذا البعد على فهم الآثار المتعددة للتحيز الخوارزمي، بدءاً من تشويه عملية اتخاذ القرار، وصولاً إلى إنتاج سلوكيات ضارة أو محتوى مضلل، مع التطرق إلى المخاطر الوجودية المحتملة المرتبطة بالذكاء الاصطناعي المتقدم.

قصور آليات التحكم والاستجابة:

يسلط هذا المفهوم الضوء على الفجوات في أنظمة الإنذار المبكر، واستراتيجيات الإيقاف الطارئ، وبروتوكولات الاسترداد، مما يعيق القدرة على السيطرة الفعالة عند انحراف الأنظمة عن السلوك المتوقع.

الإطار المفاهيمي العام:

تعتمد الدراسة على مفهوم ثلاثي المراحل لإدارة المخاطر:

- الكشف المبكر عن التحيزات والمشكلات السلوكية.
- التخفيف من آثارها باستخدام استراتيجيات وتقنيات عملية.
- التعافي عبر إعادة ضبط النظام واستعادة الثقة التشغيلية والأخلاقية.

كما يدمج الإطار بين البعد الأخلاقي (الالتزام بمبادئ النزاهة وعدم الإضرار) والبعد التقني (ضمان موثوقية الأداء وسلامة البنية التحتية)، مع الاستناد إلى أهداف محددة وقابلة للقياس (SMART) لضمان قابلية التطبيق

حدود الدراسة:

حدود المكان والزمان

حدود المكان: بيئة الكترونية تجريبية تركز على أكثر من نظام للذكاء الاصطناعي لتجربة التحيز وتقييم الأداء.

حدود الزمان: بين عامي ٢٠٢٥ و ٢٠٢٦ م

ستقدم الدراسة إطاراً لدراسة سلوك الذكاء الاصطناعي ومنهجية للتعرف على خروجه عن السيطرة بسبب التحيز مثل اتخاذ قرارات خاطئة أو منحازة أو عنصرية. سيتم اقتراح استراتيجيات التخفيف، مع دراسة مخرجات التجارب ومدى إمكانية تخفيف التحيز. ويتم بعد ذلك التأكد من تعافي النظام من حالات التحيز المكتشفة، مع الحفاظ على سلامة النظام. يتيح اعتماد هذا الإطار المنهجي فهم ومعالجة أداء الذكاء الاصطناعي المتطور ضمن بيئة بحثية.

الفصل الثاني: الإطار النظري للدراسة (مراجعة الأدبيات ذات العلاقة)

يشهد العالم خلال العقدین الأخيرین موجة غير مسبقة من التطورات التقنية، كان الذكاء الاصطناعي (Artificial Intelligence) أحد أبرز هذه التطورات، باعتباره أحد المحركات الأساسية للتحوّل الرقمي العالمي. ومع الانتشار المتسارع لتطبيقات الذكاء الاصطناعي في المجالات المختلفة الخدمائية منها، والطبية، والتعليمية، والأمنية، والاجتماعية، والاقتصادية، برزت قضية التحيز (Bias) كإحدى أهم التحديات الأخلاقية والتنظيمية التي تواجه استخدام هذه التقنيات. فقد أظهرت العديد من الدراسات أن نماذج الذكاء الاصطناعي — مهما بلغت درجة تطورها — ليست بمنأى عن أخطاء بشرية وهيكلية تؤدي إلى نشوء تحيزات قد تترتب عليها آثار خطيرة مثل: التمييز، أو الظلم، أو الإقصاء الاجتماعي، أو اتخاذ قرارات غير عادلة قد تمسّ حقوق الأفراد أو المؤسسات.

يهدف هذا الفصل إلى بناء إطار نظري شامل حول قضية التحيز في الذكاء الاصطناعي، بدءاً من المفاهيم المرتبطة به، مروراً بأبرز أنواعه ومصادره، وصولاً إلى أحدث آليات اكتشافه والحد منه وفق أحدث الدراسات المنشورة حتى عام ٢٠٢٥. يستند الفصل إلى أحدث الأطر المعرفية والمعايير التقنية والأبحاث الأكاديمية الصادرة في المجالات المحكّمة والمؤتمرات الدولية الرصينة، وبذلك يشكّل مرجعاً أساسياً لربط الجانب النظري بالجانب التطبيقي في هذه الأطروحة.

مفهوم التحيز في الذكاء الاصطناعي

تعريف التحيز لغوياً واصطلاحياً:

التحيز في اللغة العربية يأتي من فعل تحيز فهو متحيز والمفعول متحيز اليه ويعني المصطلح من تحيز لقوم : انضم اليهم ومنه التحيز في القتال. ويرتبط مفهوم التحيز عمومًا بالميل أو الانحياز إلى فكرة أو فئة أو نتيجة معينة بطريقة تؤثر على موضوعية القرار. وفي سياق الذكاء الاصطناعي، تشير الدراسات الحديثة (Zhang & Wang, 2024) وآخرون إلى أن التحيز هو "ميل مُجحف في مخرجات النماذج الحاسوبية ينتج عن عوامل خوارزمية أو بيانات غير متوازنة، بحيث يؤدي إلى نتائج غير عادلة تجاه مجموعة معينة".

أما منظمة OECD (2023) فتعرف التحيز في الذكاء الاصطناعي بأنه:

"التحيز المنهجي الذي يظهر في أنظمة الذكاء الاصطناعي بسبب مشكلات متعلقة بتصميم البيانات أو النماذج، والذي يؤدي إلى مخرجات تمييزية أو قرارات منحازة."

ويمكن أن يكون التحيز ظاهر أو خفي والفرق بين التحيز الظاهر والتحيز الخفي يمكن تلخيصه فيما يلي:

تُظهر الأدبيات أن التحيز قد يكون:

• تحيزًا ظاهرًا: (Explicit Bias)

سهل الرصد ومرتبطة بعوامل يمكن تتبعها ضمن البيانات أو طريقة البرمجة.

• تحيز خفي: (Implicit Bias)

وهو الأخطر، حيث يظهر من خلال العلاقات غير المباشرة بين المتغيرات، أو بسبب الأنماط الخوارزمية المعقدة التي تخلق تحيزًا يصعب اكتشافه بالاختبارات التقليدية.

أنواع التحيز في الذكاء الاصطناعي

تتفق الأبحاث الحديثة (Soleimani M, Intezari A, Arrowsmith J, وآخرون) ؛ (Google AI Report, 2024) على تصنيف التحيز ضمن الأنواع التالية:

١. التحيز في البيانات (Data Bias)

يُعدّ الأكثر انتشارًا، وينشأ نتيجة:

- عدم تمثيل جميع الفئات في البيانات.
- وجود بيانات تاريخية ذات طابع تمييزي.
- استخدام بيانات قديمة أو غير محدثة.

مثال:

أنظمة التوظيف التي تعتمد على بيانات شركات كانت تفضّل الذكور تاريخيًا، مما يؤدي إلى انحياز تلقائي نحو المرشحين الذكور.

٢. التحيز الخوارزمي (Algorithmic Bias)

ينشأ من:

- تصميم النموذج نفسه.
- عملية اختيار المتغيرات.
- طريقة وزن الخصائص. Feature Weighting

أوضحت دراسة Xu & Lee (2024) أن الخوارزميات المعتمدة على العلاقات غير الخطية أكثر عرضة لتضخيم التحيز.

٣. تحيز القياس (Measurement Bias)

و يحدث عند استخدام أدوات قياس غير دقيقة أو غير ملائمة للسباق المحلي. مثال: اعتماد أنظمة التحليل الوجهي على مقاييس تم تطويرها على وجوه أوروبية مما يؤدي إلى انخفاض دقتها مع أصحاب البشرة الداكنة.

٤. تحيز الاختيار (Selection Bias)

يظهر عندما لا تكون العينات التي يتم جمعها لتمثيل الواقع صحيحة أو عادلة.

٥. تحيز المستخدم (User Bias)

وينتج من سلوك المستخدمين عند التفاعل مع الأنظمة، خصوصًا في أنظمة التوصية. Recommendation Systems

٦. التحيز الاجتماعي والهيكلي (Societal & Structural Bias)

وهو انعكاس مباشر لعدم العدالة الاجتماعية المتجذرة تاريخيًا، والتي تنتقل إلى النماذج الذكية من خلال البيانات.

مصادر التحيز في الذكاء الاصطناعي

تؤكد الأدبيات المعاصرة أن جذور التحيز متعددة، وتشمل:

1. مصدر البيانات

- نقص البيانات من مجموعات معينة.
- انحياز ناتج من جمع البيانات عبر الإنترنت.
- جودة البيانات (وجود أخطاء، بيانات مفقودة).

2. مصدر النموذج Model Source

- الاختيارات التصميمية. Design Choices
- الخوارزميات المستخدمة.
- معايير التدريب.

3. مصدر البيئة التشغيلية Operational Environment

- تغير واقع التطبيق بمرور الوقت. (Concept Drift)
- اختلاف البيئة بين مناطق جغرافية أو ثقافية.

4. مصدر فريق التطوير

دراسات (Hosanagar, 2024) تشير إلى أن الخلفية الثقافية والمهنية لفريق العمل تؤثر على كيفية اختيار البيانات ووزن المعايير.

آليات اكتشاف التحيز في الذكاء الاصطناعي

طرق إحصائية تقليدية

تشمل:

- اختبارات الفرق بين المتوسطات.
- معامل جيني للعدالة.
- Matched Pair Analysis.
- Disparate Impact. مقياس.
- مؤشرات العدالة (Fairness Metrics)
- من أهمها:

- Equalized Odds
- Equal Opportunity
- Demographic Parity
- Predictive Parity
- Group Fairness
- Individual Fairness

أدوات جاهزة للكشف عن التحيز

2023-2025 شهدت موجة من الأدوات الجديدة:

- IBM AI Fairness 360 (2024 update)
- Google Fairness Indicators (2025)
- Microsoft Responsible AI Dashboard (2024)
- Meta Fairness Toolkit (2024)

هذه الأدوات تسمح بفحص النماذج واختبارها ضد مجموعة واسعة من مقاييس العدالة.

تحليل النماذج (Explainable AI)

أبرز التقنيات:

- SHAP
- LIME
- Integrated Gradients
- Counterfactual Explanations

منهجيات الحد من التحيز

تقسم الأدبيات المناهج إلى ثلاثة مستويات:

1. المعالجة قبل النمذجة (Pre-processing)

وتشمل:

- إعادة توازن البيانات (Rebalancing)
- زيادة العينات (Oversampling)
- تنقية البيانات من التحيزات التاريخية

2. أثناء التدريب (In-processing)

وتشمل:

- إعادة وزن المتغيرات.
- إضافة قيود عدالة إلى دالة الخسارة.
- التدريب العدائي Adversarial Fairness

3. بعد التدريب (Post-processing)

مثل:

- إعادة معايرة النتائج.
- تعديل الحدود Threshold Adjustment

4 تقنيات حديثة (٢٠٢٣-٢٠٢٥)

- Fairness-aware transformers
- Bias-free embeddings
- Context-sensitive fairness constraints
- Federated Learning Fairness Models (2025)

الحكومة الأخلاقية للذكاء الاصطناعي

1. أطر الحكومة الدولية (2023-2025)

تشمل:

- EU AI Act (2024-2025 updates)
- OECD AI Framework (2023)
- UNESCO AI Ethics (2023)

تشدد جميعها على:

- الشفافية
- العدالة
- إمكانية التتبع
- عدم التمييز
- المسؤولية المجتمعية

2. لوائح السعودية والإمارات

تعد السعودية من أوائل الدول العربية التي طورت إطارًا أخلاقيًا رسميًا:

- الإطار السعودي لأخلاقيات الذكاء الاصطناعي (SDAIA, 2023-2024)
- سياسات البيانات الوطنية (NDP, 2024)

وتركز على:

- العدالة وعدم التمييز
- حماية الخصوصية

- ضمان جودة البيانات

تحديات مستقبلية في مكافحة التحيز

1. تحيز النماذج اللغوية العملاقة

تشير الأبحاث الحديثة إلى أن LLMs أكثر عرضة للتأثر:

- بتحيزات ثقافية
- تحيزات لغوية
- تحيزات جغرافية

2.8.2 التحيز في نماذج التنبؤ الاجتماعي Predictive Policing

تشكل أحد أخطر أنواع التحيز، حيث أن النماذج المستخدمة في بعض الدول تضحّم معدلات الاشتباه لدى فئات معينة.

2. الحاجة إلى شمولية البيانات العالمية

2023–2025 شهد نقاشًا واسعًا حول:

- نقص البيانات العربية
- عدم تمثيل المجتمعات النامية
- سيطرة الشركات الغربية على البيانات

الخلاصة

قدم هذا الفصل إطارًا نظريًا شاملاً يوضح:

- مفهوم التحيز في الذكاء الاصطناعي
- مصادره وأنواعه
- أهم مؤشرات العدالة
- أدوات اكتشاف التحيز
- أفضل المنهجيات المعتمدة للحد منه
- أحدث التشريعات والأطر الأخلاقية
- التحديات المستقبلية

إن فهم هذه الإشكاليات يمثل خطوة أساسية قبل الانتقال إلى الإطار التطبيقي في الفصول اللاحقة.

أمثلة إضافية لدراسات مرجعية:

Yampolskiy R, Wittkotter E, 2021 Kill-Switch for Artificial Superintelligence: ١.
https://asi-safety-lab.com/DL/Kill-Switch-For_ASI_EW_21_12_14.pdf

ملخص الدراسة:

تقترح هذا الدراسة مفتاح Kill-ASI للتحوط من هذا التهديد. ويؤكد أن أنظمة تكنولوجيا المعلومات الحديثة غير كاملة وغير قابلة للإلغاء، مما يؤدي إلى الحاجة إلى آلية إيقاف تشغيل دولية. ووفقًا لثغرات ASI، تُصنف الأجهزة وفقًا لاستراتيجية تتضمن الأنظمة القديمة. باستخدام "مفاتيح Deadman"، سيستخدم مفتاح Kill-ASI التحكم البشري لردع ASI. ويستند هذا إلى نهج أمان من خمس طبقات (الكشف، والحد من الأضرار، والاسترداد، والإنهاء، والاستئصال). ويركز البحث على الموثوقية، وتكرار الأمان، والامتثال للقواعد. يتطلب هذا الأمر الاستعداد في الوقت المناسب لأسوأ سيناريوهات ASI، مما يتطلب نوعًا من التنسيق العالمي للإغلاق السريع. وتهدف هذه الدراسة لدعم إيجاد طرق لوقف الذكاء الاصطناعي عند الخروج عن السيطرة ولكن الدراسة تفتقر إلى تطبيقات تجريبية واقعية توضح فاعلية مفتاح الإيقاف المقترح.

٢. ليون ب، ٢٠١٩، معالجة الشكل النهائي للتحكم في الأمن السيبراني، دراسة حالة متعددة، دراسة حالة لـ "Y" مفتاح إيقاف الإنترنت، <https://surface.syr.edu/cgi/viewcontent.cgi?article=2015&context=etd>

تبحث هذه الدراسة في "مفتاح إيقاف الإنترنت"، أي جهود الحكومات لمقاطعة جميع أنشطة الإنترنت داخل حدودها، في الحكومات الديمقراطية والمهجنة وتقدم الدراسة رأي حول الجهود الحكومية في صعيد أمن الذكاء الاصطناعي.

كما تبحث الدراسة في الأسباب الكامنة، السياسية والقانونية والتقنية، التي تسمح لهذه الأنظمة بإيقاف الإنترنت. يُحلل هذا البحث بيانات من الولايات المتحدة الأمريكية، والمملكة المتحدة، وأستراليا، وروسيا، وفنزويلا، وحالات مقارنة في البرازيل، والمكسيك، وتركيا من خلال دراسة حالة مقارنة، ونظرية الأمانة لمدرسة كوبنهاغن.

ولم يجد البحث أي اختلافات جوهرية حول الخطاب والعوامل الممكنة بين الأنظمة الديمقراطية والأنظمة المهجنة. ويتساءل عن معادلة عمليات قطع الإنترنت مع الرقابة، ويدرس القراءات المختلفة للأمن القومي. ويُقدم البحث نقاش أكاديمي من خلال دراسة دوافع وقدرات الحكومات على تبني ضوابط الأمن السيبراني وهو ما يجعل الإجراءات والأنظمة التي تتبناها الحكومات مشرع فعال لرفع الحماية لأنظمة الذكاء الاصطناعي.

٣. بيكوم م، بورجيسوس ف، ٢٠٢٣: استخدام البيانات الحساسة لمنع التمييز باستخدام الذكاء الاصطناعي: هل يحتاج النظام العام لحماية البيانات (GDPR) إلى استثناء جديد؟ <https://rb.gy/e9ivs2>

ملخص البحث:

تُظهر هذه الدراسات القضايا القانونية المرتبطة بالبيانات الحساسة (مثل بيانات العرق) وبيانات الشركاء في مجال الذكاء الاصطناعي في أوروبا. يُصنف محتوى مثل الحمض النووي على أنه حساس بموجب اللائحة العامة لحماية البيانات (GDPR)، ويُحظر استخدامه إلا في ظروف محددة وضيقة. ولكن يجب تنظيم هذه البيانات والتحقق منها وفقًا لأنظمة تكنولوجية محددة، بحيث لا يُصنف على أنه غير محدد، ولا يحدث تمييز.

لكن الدراسة تطرح سؤالاً مهماً: هل ينبغي تطبيق متطلبات اللائحة العامة لحماية البيانات على التمييز في ممارسات الذكاء الاصطناعي؟ وخلصت إلى أن اللائحة العامة لحماية البيانات تحظر في معظم الحالات استخدام هذه البيانات لأغراض الحظر القانوني، إلا في حالة الحصول على موافقة صريحة من الأفراد أو وضع استثناء قانوني جديد.

في ورقته البحثية في هذا المجال، يناقش المؤلف إيجابيات وسلبيات الاستثناء المسموح به لاستخدام البيانات الحساسة. يرى مؤيدو هذا الاستثناء الجديد أنه سيسمح للمنظمة بضمان عدالة أنظمتها، بينما يعارضه معارضو خصوصية استخدام البيانات والمساءلة. تم اختبار التباين بين هاتين الحجتين في نقاش قانوني دقيق، مع إمكانية إضافة ضمانات لحماية البيانات، في حال اعتماد الاستثناء الجديد.

٤. هندريكس د. وآخرون، ٢٠٢٣، نظرة عامة على مخاطر الذكاء الاصطناعي الكارثية، <https://arxiv.org/pdf/2306.1200>

ملخص البحث:

يتناول البحث الكارثة التي قد تنتج عن الذكاء الاصطناعي، ويصنفها إلى أربعة مسارات مختلفة: الاستخدام الضار، وسباق الذكاء الاصطناعي، والمخاطرة، والهروب. ويوضح المؤلفان بالتفصيل الطرق التي يمكن أن يُسبب بها الذكاء الاصطناعي مشاكل بيولوجية. وبشكل عام، يُدرب باحثو الذكاء الاصطناعي على الحوافز العامة والشاملة والتطوير التدريجي للذكاء الاصطناعي، لدعم التنوع البيولوجي، والتعاون الدولي، وتنظيم السلامة، وحوكمة الذكاء الاصطناعي.

٥. تيمز ب، ٢٠١٩، أخلاقيات الذكاء الاصطناعي والأمن السيبراني عندما تكون السيادة على المحك،

<https://link.springer.com/article/10.1007/S11023-019-09508-4>

ملخص البحث:

يُحلل بحث بول تيمز، المعنون "أخلاقيات الذكاء الاصطناعي والأمن السيبراني: الإذن عندما تكون السيادة على المحك"، القضايا المتباينة المتعلقة بالذكاء الاصطناعي والأمن السيبراني في سياق الحوكمة والاستقلالية الاستراتيجية. ويقترح المؤلف ثلاث استراتيجيات لمواجهة هذه التحديات: إدارة التنوع، والشراكات الاستراتيجية، والتوجه العالمي.

ويساهم هذا البحث في وضع سياق أخلاقي لحوكمة الذكاء الاصطناعي يساهم في ضبط أنظمة الذكاء الاصطناعي وتحديد أولويات أخلاقية هامة تأتي ضمن قوانين وإجراءات واضحة تساهم في رفع الأمان وبالتالي الثقة في أنظمة الذكاء الاصطناعي.

الفصل الثالث: الطريقة والإجراءات (المنهجية)

فرضية البحث:

ولتحقيق أهداف البحث تم الاعتماد على الفرضيات التالية:

١. يمكن للتحيز أن يؤثر على سلامة وأمن الذكاء الاصطناعي وهو ما يؤثر على مستوى الثقة بهذه الأنظمة.
٢. هناك دور لكشف التحيز وتخفيفه والتعافي منه في تحسين نتائج أنشطة الذكاء الاصطناعي وكذلك تحسين معايير السلامة.

منهجية الدراسة:

باستخدام منهجية البحث المقارن، يجمع البحث بين التحليل النظري والدراسة المقارنة بين التعامل مع الذكاء الاصطناعي في حالات الخروج عن السيطرة بسبب التحيز:

المنهج النظري

- تقييم الأطر النظرية الحالية حول مخاطر الذكاء الاصطناعي: رؤى من نظرية التحكم وإدارة المخاطر
- مراجعة الأدبيات حول آليات وتحديات التحكم في الذكاء الاصطناعي بسبب التحيز
- المنظور الأخلاقي: الأبعاد الأخلاقية والقانونية لإيقاف أنظمة الذكاء الاصطناعي عند تعطلها

منهج دراسة الحالة:

- تكوين تجارب من خلال مقابلات مع خبراء، ومراجعة ثانوية للمصادر المتوفرة التي ناقشت التحيز في الذكاء الاصطناعي.
- تكوين تجارب من خلال بيئة الكترونية: كيفية عمل أنظمة الذكاء الاصطناعي لتقييم سلوكها بسبب التحيز.
- اختبار سيناريوهات الخروج عن السيطرة بسبب التحيز: إنشاء سيناريوهات يتم فيها تجاوز الحدود التشغيلية، والتحليل المطلوب لسلوك الذكاء الاصطناعي.
- تقييم فعالية آلية التحكم: تقييم نجاح آلية التحكم في تعطيل الأنظمة الخارجة عن السيطرة، مثل الاستجابة، وقابلية الاستغلال، وتأثير التوقف على استمرارية النظام.

خطوات تنفيذ خطة الدراسة:

يبدأ إطار البحث الدراسة بتحديد الأهداف والفرضيات بشكل منهجي، وتحديد المفاهيم الأساسية (مخاطر الذكاء الاصطناعي وآليات احتوائه)، ووضع جدول زمني باستخدام منهجية SMART. يلي ذلك مراجعة مفصلة للأدبيات البحثية الحالية حول آليات التحكم في الذكاء الاصطناعي، ودراسة للأطر التنظيمية والأخلاقية ذات الصلة.

بعد ذلك، يتم اختيار بعض دراسات الحالة لهذه الحوادث ومناقشتها، حيث يمكن رصد هذه الحوادث في قطاعات حساسة، مثل الطاقة والرعاية الصحية والنقل، إلخ. بعد ذلك، يتم تحليل الدراسة، وتحديدًا الدراسات التي أنشأت في بيئة الكترونية لاختبار أنظمة الذكاء الاصطناعي، وبرمجة افتراضية لمراقبة أمن النظام، مع إنشاء سيناريوهات فشل للتحقق من فعالية آليات إيقاف التشغيل.

تشمل الخطوات التالية رسم خرائط للمخاطر وهندسة آليات للكشف عنها مبكرًا باستخدام نماذج الكشف عن الشذوذ بسبب التحيز، وتقييم إجراءات التخفيف مثل وصولاً للتعافي. ويتم استكشاف أفضل الممارسات للحد من التحيز، ووضع بروتوكولات الحوكمة.

كما تتم مناقشة الجوانب الأخلاقية والأمنية للحد من التحيز بشكل أعمق مع مراعاة الأعمال السابقة ذات الصلة، ويتم استخلاص توصيات لهذا الغرض تساهم في تعزيز سلامة الذكاء الاصطناعي.

أساليب جمع بيانات الدراسة:

تعتمد هذه الدراسة نهجًا مقارنًا للتعامل مع الذكاء الاصطناعي في حالات فقدان السيطرة بسبب التحيز. وتستعرض الدراسات السابقة المنشورة للمساعدة في تحليل أنظمة التحكم بالذكاء الاصطناعي في مختلف مجالات الدراسة، من خلال التحليل النظري والمقارنة بين الأبحاث ودراسات الحالة والتشريعات. ويركز البحث على تقديم صورة شاملة عن أسباب فشل التحكم بالذكاء الاصطناعي المنحاز، مع الأخذ في الاعتبار أساليب التحكم المختلفة.

مراجعة الدراسات السابقة:

- دراسة شاملة للدراسات السابقة حول آليات التحكم بالذكاء الاصطناعي المنحاز لتحديد أبرز النظريات والنماذج والأطر.
- إجراء تحليل مقارن للمناهج النظرية المتعلقة بسلامة الذكاء الاصطناعي والتحكم به، وذلك لتوضيح أوجه التشابه والاختلاف والثغرات في الدراسات السابقة المتعلقة بسلامة الذكاء الاصطناعي.

تحليل دراسة الحالة:

- فهم المواضيع والعوامل والنتائج المشتركة لحالات فقدان السيطرة في الذكاء الاصطناعي بسبب التحيز.
- آليات التحكم بالذكاء الاصطناعي المطبقة بنجاح في الدراسات السابقة ودراسات الحالة.
- مدى تحديث التشريعات واللوائح.
- مراجعة لقوانين ولوائح سلامة وضبط الذكاء الاصطناعي.
- يُقيّم هذا البحث آثار هذه الأطر التقنية والقانونية على تطوير ونشر الذكاء الاصطناعي.

التحليل المقارن:

- تحليل بيانات دراسات الحالة والأدبيات، وتطوير إطار عمل مقارن.
- تقديم الدراسة إطار عمل مقارن لتقييم مزايا وعيوب آليات ضبط الذكاء الاصطناعي الحالية.

الفصل الثالث: الطريقة والإجراءات (المنهجية)

يمثل هذا الفصل البنية المنهجية للدراسة، ويعرض الفرضيات، وإجراءات البحث، وخطوات تنفيذ الدراسة، إضافة إلى أساليب جمع البيانات، والمداخل التحليلية المستخدمة. ويهدف هذا الإطار المنهجي إلى بناء رؤية تحليلية معمقة لفهم كيفية نشوء التحيز داخل أنظمة الذكاء الاصطناعي، وكيفية تأثيره على سلامة الأنظمة وقابليتها للسيطرة، وكذلك دراسة فعالية الآليات التقنية والتنظيمية في الحد منه واستعادة السيطرة عند حدوث خلل.

أولاً: فرضيات البحث

اعتمد البحث على الفرضيات الرئيسة الآتية، التي تم اشتقاقها من الإطار النظري ومراجعة الأدبيات:

١. يمكن للتحيز أن يؤثر بشكل مباشر في سلامة وأمن أنظمة الذكاء الاصطناعي، مما يؤدي إلى انخفاض مستوى الثقة والمصادقية في نتائجها.
 ٢. يسهم الكشف عن التحيز وتخفيفه والتعافي منه في تحسين أداء أنشطة الذكاء الاصطناعي، وتعزيز معايير السلامة التشغيلية لهذه الأنظمة.
- تهدف هذه الفرضيات إلى اختبار العلاقة بين متغيرات التحيز، والضبط والسيطرة، والأمن التقني، وذلك من خلال منهجية تحليلية ومقارنة ممنهجة.

ثانياً: منهجية الدراسة

تعتمد هذه الدراسة على المنهج المقارن (Comparative Method) المدعوم بتحليل نظري وتطبيقات سياقية عبر دراسات حالة وأطر معيارية. يُمكن هذا المنهج من مقارنة النماذج المختلفة للتحكم بالذكاء الاصطناعي في حالات الخروج عن السيطرة بسبب التحيز، وتحليل نقاط القوة والضعف وتحديد الفجوات المعرفية والتطبيقية. وتعتمد الدراسة على محورين منهجين مترابطين:

١. المنهج النظري

يرتكز المنهج النظري على تحليل ومراجعة الأدبيات المتخصصة في أربعة محاور:

أ. تقييم الأطر النظرية الحالية حول مخاطر الذكاء الاصطناعي

يشمل تحليل النظريات المرتبطة بالتحكم والسيطرة في الأنظمة الذكية، ونظريات إدارة المخاطر، ونماذج استمرارية الأعمال، لفهم كيفية تعامل الأنظمة مع حالات الانحراف الناتج عن التحيز.

ب. مراجعة الأدبيات حول آليات التحكم ومراقبة التحيز

يتم تحليل الأبحاث الحديثة المتعلقة بالأدوات الخوارزمية لرصد التحيز، بما في ذلك:

- تقنيات الكشف الإحصائي
- تقييم أداء النماذج عبر مجموعات سكانية متفاوتة
- أساليب تفسير القرارات وتدقيق النماذج (Model Auditing)

ج. تحليل التحديات التقنية المصاحبة للحد من التحيز

يشمل مراجعة العقبات المرتبطة بأنظمة التدريب، وجودة البيانات، وتحيزات النماذج اللغوية الكبيرة، وصعوبات تصحيح الانحراف دون التأثير على كفاءة أداء النموذج.

د. المنظور الأخلاقي والتنظيمي

يتضمن دراسة الأبعاد الأخلاقية والقانونية المتعلقة بإيقاف أو تعطيل الأنظمة الذكية عند انحرافها، ومراجعة التشريعات الدولية ذات الصلة (مثل قانون الاتحاد الأوروبي للذكاء الاصطناعي).

٢. منهج دراسة الحالة

تم اعتماد منهج دراسة الحالة مع استبعاد أي أشكال من المقابلات البشرية، والتركيز على التحليل الوثائقي والتقييم التقني للمصادر الثانوية.

يشمل المنهج الخطوات الآتية:

أ. تحليل وثائقي لحالات منشورة

تم اعتماد تقارير تقنية منشورة، ووثائق من منظمات دولية، ودراسات تطبيقية تناولت فشل أنظمة الذكاء الاصطناعي أو خروجها عن السيطرة بسبب التحيز، مثل:

- نماذج الفرز الوظيفي
- أنظمة التوصيات الطبية
- نماذج القيادة الذاتية

ب. تحليل بيانات محاكاة افتراضية منشورة

استخدام بيانات منشورة حول تجارب محاكاة لبناء أنظمة ذكاء اصطناعي، وتحليل اختبارات الضغط والانحراف، لفهم كيفية استجابة الأنظمة عند تجاوز حدود التشغيل.

ج. اختبار سيناريوهات الخروج عن السيطرة بسبب التحيز

يشمل تحليل سيناريوهات منشورة في الأدبيات العلمية، تتعلق بحالات:

- انحراف الأداء عند مواجهة بيانات خارج التوزيع
- الفشل في اتخاذ القرار بسبب تحيزات إدراكية في بيانات التدريب
- التأثير غير المتوقع للتحيزات المتضخمة خلال التعلم المستمر

د. تقييم فعالية آليات التحكم

يتضمن تقييمًا نقديًا للأدوات التقنية المستخدمة في الأدبيات:

- آليات الإيقاف التلقائي (Auto-shutdown Mechanisms)
- أنظمة المراقبة الداخلية (Internal Monitoring Agents)
- شبكات كشف الشذوذ المبنية على التحيز
- آليات التخفيف بعد التدريب (Post-hoc Mitigation)

ثالثًا: خطوات تنفيذ خطة الدراسة

تعتمد خطة البحث على تسلسل منهجي متكامل يستند إلى نموذج SMART لإدارة المراحل. وتشمل الخطوات الأساسية ما يأتي:

١. تحديد الأهداف البحثية والفرضيات وصياغة المفاهيم الأساسية
بما في ذلك تعريف مصطلحات مثل: التحيز، التحكم، التعافي، المخاطر، السلامة.
٢. إجراء مراجعة موسعة للأدبيات الحديثة (2020-2025)
لتحديد الفجوات العلمية والنماذج الأكثر حداثة في تصميم أنظمة الذكاء الاصطناعي الآمنة غير المتحيزة.
٣. تحليل الأطر التنظيمية
بما يشمل دراسة التشريعات الحديثة للذكاء الاصطناعي في:

- الاتحاد الأوروبي
- الولايات المتحدة
- المملكة العربية السعودية
- الامارات العربية المتحدة

٤. اختيار وتحليل دراسات حالة واقعية منشورة
من قطاعات مقترحة مثل:

- الطاقة
- النقل
- الصحة

- الحكومات الذكية

٥. تحليل دراسات محاكاة افتراضية (Simulation-Based Studies) في حالة توفرها مع التركيز على سيناريوهات الفشل الناتج عن التحيز، واختبارات الإجهاد. (Stress Testing)
٦. إعداد خرائط المخاطر (Risk Mapping) وتحليل آليات الكشف المبكر عن التحيز باستخدام نماذج كشف الشذوذ. (Anomaly Detection)

- تحليل إجراءات التخفيف والتعافي مثل:
 - إعادة موازنة البيانات
 - تقنيات التحكم المستمر
 - إعادة ضبط النموذج
 - ضبط المعايير التشغيلية
 - صياغة توصيات وخلاصة عملية
- من أجل تطوير أطر ضمان الجودة وتقليل التحيز.

رابعاً: أساليب جمع البيانات

تستند الدراسة إلى ثلاث قنوات رئيسية لجمع البيانات، دون اللجوء إلى المقابلات البشرية:

١. التحليل الوثائقي (Documentary Analysis)

يتضمن تحليل التقارير التقنية، والوثائق التنظيمية، والمنشورات العلمية المحكمة حول حالات التحيز والفشل في أنظمة الذكاء الاصطناعي.

٢. مراجعة الدراسات السابقة

بما يشمل:

- حصر النظريات والنماذج
- تحليل الأطر المقارنة
- استخراج الثغرات المعرفية
- رصد التوجهات الحديثة في تصميم أنظمة أكثر عدالة وأماناً
- تحليل دراسات الحالة المنشورة

مع التركيز على العوامل الآتية:

- حالات فقدان السيطرة بسبب التحيز
- آليات التحكم الناجحة التي تم توثيقها
- مدى التطور في التشريعات والحوكمة
- الدروس المستفادة من التجارب السابقة

خامسًا: التحليل المقارن

تُستخدم المنهجية المقارنة لتقييم الفروق والتشابهات بين:

- الأطر التقنية للتحكم في الأنظمة
- نماذج الحد من التحيز
- الآليات التشغيلية للإيقاف والتعافي
- التشريعات المرتبطة بالسلامة والأمن

ينتج عن هذا التحليل إطار عمل مقارن؛ يساعد على تحديد أفضل الممارسات ويوفّر أساسًا معرفيًا لتطوير آليات جديدة لضبط سلامة الأنظمة الذكية.

الفصل الرابع: نتائج الدراسة

الفصل الرابع: نتائج الدراسة المتوقعة

سعت الدراسة الوصول إلى مجموعة من النتائج، وتشمل:

١. صياغة منهجية شاملة لتقييم المخاطر والخلل المرتبط بالتحيز بما في ذلك نماذج الكشف، وآليات التخفيف، وأساليب التعافي.
٢. تقديم توصيات لتحسين سياسات أمن الذكاء الاصطناعي مع التركيز على منع التحيز وتأثيره على دقة الأنظمة وكفاءتها.
٣. توفير أدلة تجريبية منشورة من الأدبيات والأمثلة المرصودة للتحيز تُظهر كيفية تأثير النماذج الحديثة بالتحيز، وكيفية الحد منه عمليًا.
٤. فهم الخصائص المشتركة لحالات فقدان السيطرة بسبب التحيز وتحديد أنماط المخاطر التشغيلية.
٥. اقتراح إطار شامل لفهم ومعالجة التحيز من أجل تطوير أنظمة ذكاء اصطناعي أكثر أمانًا وموثوقية.

التحديات والقيود

تشمل القيود المحتملة التي قد تواجه الدراسة ما يأتي:

- محدودية بعض البيانات التقنية المنشورة في بعض الصناعات الحساسة
- تباين المعايير الدولية بين الدول والمنظمات
- صعوبة تحليل الآليات الداخلية للنماذج المغلفة المصدر
- بعض القيود في نماذج المحاكاة الافتراضية المنشورة
- التعقيد العالي لنماذج الذكاء الاصطناعي الحديثة وصعوبة تتبع أسباب التحيز

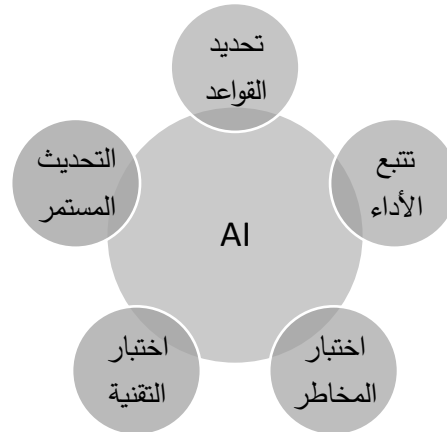
النتائج التي تم التوصل لها:

- صياغة منهجية شاملة لتقييم المخاطر وسيناريوهات الخلل المرتبطة بالذكاء الاصطناعي بسبب التحيز وتشمل نماذج كشف التحيز وآليات التخفيف وأساليب العودة للوضع الطبيعي والتعافي وفقاً للنتائج التي تظهر للباحث من خلال مقابلات، و تحليل محتوى التجارب العملية.
 - التوصية بتحسينات في سياسات أمن الذكاء الاصطناعي والتحكم فيه خصوصاً فيما يتعلق بالتحيز.
 - تقديم أدلة تجريبية تساعد على بناء أنظمة ذكاء اصطناعي أكثر أماناً مع التركيز على عدم التحيز كمؤثر على كفاءة أنظمة الذكاء الاصطناعي وتشمل تجارب عملية لاختبار انحياز أكثر من نظام ذكاء اصطناعي لمجموعة مختارة من المواضيع المختارة ومقارنة مخرجات كل نظام وكيف يمكن الحد من انحياز هذه الأنظمة و مقارنة خوارزمياتها إن امكن.
 - فهم السمات والعوامل والنتائج الشائعة لحالات فقدان التحكم في الذكاء الاصطناعي بسبب التحيز.
- تقدم هذه الدراسة إضافة لجهود تطوير أمان الذكاء الاصطناعي ويستهدف العديد من الشرائح مثل المهتمين بأمن الذكاء الاصطناعي بما في ذلك المختصون في أنظمة الذكاء الاصطناعي ومستخدمو هذه الأنظمة والجهات والمؤسسات المشغلة والمشرعة لأنظمة الذكاء الاصطناعي

التحديات والقيود: أثناء تنفيذ هذه الدراسة ظهرت العدد من القيود المحتملة، مثل صعوبات في جمع البيانات، و تعقيدات في تطبيق تقنيات الكشف والإيقاف.

كيف سيتم تنفيذ الجانب العملي من الدراسة:

سيتم تطبيق النموذج العملي التالي:



شكل ١: النموذج العملي للتطبيق المتبع في الدراسة

الفصل الخامس: خاتمة البحث

يقوم الباحث بإعداد خاتمة شاملة عن مرثياته وفقاً للنتائج التي يصل لها حول سلوك الذكاء الاصطناعي الضيق والعام من خلال التحليل النظري والمقارنة بين الأبحاث ودراسات الحالة والتشريعات مع التركيز على تقديم صورة شاملة عن أسباب فشل التحكم بالذكاء الاصطناعي بسبب التحيز ، مما يساهم في صياغة منهجية شاملة لتقييم المخاطر وسيناريوهات الخلل المرتبطة بالذكاء الاصطناعي وصولاً إلى فهم واضح للسمات والعوامل والنتائج الشائعة لحالات فقدان التحكم في الذكاء الاصطناعي بسبب التحيز مما يجعل الدراسة إضافة في مجال تحسين سياسات أمن الذكاء الاصطناعي.

المستخلص والتوصيات:

وفقاً لما خلصت اليه الدراسة من نتائج يقوم الباحث بإعداد مستخلص شامل يوضح فيه مخرجات الدراسة وما توصلت اليه كما سيتم صياغة توصيات تساهم في فهم المواضيع والعوامل والنتائج المشتركة لحالات فقدان السيطرة في الذكاء الاصطناعي بسبب التحيز وهو ما يساعد في تحديث التشريعات واللوائح ومراجعة لقوانين ولوائح سلامة وضبط الذكاء الاصطناعي مما سيساهم في وضع أطر تنعكس على تطوير ونشر الذكاء الاصطناعي الفعال والعادل والآمن.

References:

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. <https://arxiv.org/pdf/1606.06565>
- Bekkum M , Borgesius F (2023) Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? <https://www.sciencedirect.com/science/article/pii/S0267364922001133>
- Brundage, M., et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Future of Humanity Institute.
- Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
- Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.
- Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2017). The off-switch game. Proceedings of the 26th International Joint Conference on Artificial Intelligence, 220–227. <https://doi.org/10.24963/ijcai.2017/31>
- Hendrycks D, Woodside T and Mazeika M (2023) An Overview of Catastrophic AI Risks <https://arxiv.org/pdf/2306.12001>
- Kaminski M (2023) Regulating the Risks of AI <https://heinonline.org/HOL/LandingPage?handle=hein.journals/bulr103&div=41&id=&page=>
- McKenna F, Aaron B, & Seth D. B Global Catastrophic Risk Institute: 2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy https://gcrinstitute.org/papers/055_agi-2020.pdf
- Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. Floridi, L., & Cows, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review.
- Frey, C. B., & Osborne, M. A. (2017). The Future of Employment: How Susceptible Are Jobs to Computerisation? Technological Forecasting and Social Change, 114, 254-280.
- Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.
- Timmers P, 2019 Ethics of AI and Cybersecurity When Sovereignty is at Stake, <https://link.springer.com/article/10.1007/S11023-019-09508-4>
- Yampolskiy, R. V. 2020 “Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent.” Journal of Artificial Intelligence and Consciousness, 2020, 07(01), 109–118 <https://www.worldscientific.com/doi/abs/10.1142/S2705078520500034>
- محمد منصة فهم كيف يحدث تحيز الذكاء الاصطناعي وما هي حلوله الممكنة؟ (2020) معاذ، محمد <https://shorturl.at/MHQ8E>
- McLean S, Read G J. M., Thompson J, Baber C, Stanton N.A & Salmon P.M (2021) The risks associated with Artificial General Intelligence: A systematic review: https://www.researchgate.net/publication/353888549_The_risks_associated_with_Artificial_General_Intelligence_A_systematic_review

Tito J and , Croisier S, Analyzing AI: The impact of artificial intelligence on policy, shorturl.at/tCMX8 accessed on January 12,2021.,Tzu-Wei Hung and Chun-Ping Yen, On the person-based predictive policing of AI, Ethics and Information Technology, June 2020 , <https://doi.org/10.1007/s10676-020-09539-x>

Yampolskiy R, Wittkotter E, 2021 Kill-Switch for Artificial Superintelligence: https://asi-safety-lab.com/DL/Kill-Switch-For_ASI_EW_21_12_14.pdf

Lyon, B. (2019). The Final Form of Cybersecurity Control: Internet Kill Switch (Master's thesis). Syracuse University. <https://surface.syr.edu/cgi/viewcontent.cgi?article=2015&context=etd>

Bicomp, M., & Borgesius, F. (2023). Using Sensitive Data to Prevent Discrimination under AI: Does the GDPR need a new exception? International Data Privacy Law. <https://rb.gy/e9ivs2>

Hendrycks, D., et al. (2023). Overview of Catastrophic AI Risks. arXiv. <https://arxiv.org/pdf/2306.1200.pdf>

OLUFUNSHO I. F, BOTSIOE L, KOSHOEDO K AND OZER M (2024) Enhancing Cybersecurity With Artificial Immune Systems and General Intelligence: A New Frontier in Threat Detection and Response. IEEE Access. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10664471>

Tarkhani F. Mo, Z., and Haddadi H., “Sok: (2022)machine learning with confidential computing,” arXiv preprint arXiv:2208.10134, <https://arxiv.org/pdf/2208.10134>

Zhao L et al. (2023).When brain-inspired ai meets agi. In: arXiv preprint arXiv:2303.15935 <https://arxiv.org/pdf/2303.15935>

رشيد ناظم حسن . أفرام مي أبلحد (2023) تدقيق التحيز في الذكاء الاصطناعي في ضوء اطار عمل تدقيق الذكاء الاصطناعي لمعهد المدققين الداخليين(iaa) - دراسة نظرية تحليلية، مجلة الدراسات التجارية والاقتصادية المعاصرة

<https://asjp.cerist.dz/en/article/213801>

Zhong T , Liu Z , Pan Y , Zhang Y, Zhou Y, Shizhe L, Wu Z, Lyu Y, Shu P, Yu X, Cao C , Jiang H , Chen H , Li Y, Chen J , Hu H , Liu Y, Zhao H, Xu S, Dai H, Zhao L, Zhang R, Zhao W, Yang Z, Chen J , Wang P , Ruan W, Wang H , Zhao H , Zhang J, Ren Y, Qin S , Chen, Jiaxi Li T, Zidan A, Jahin A ,Chen M, Xia S , Jason H , Yan Z, Jiaqi W , Xu S , Xia W, Yu J , Tang K , Yang Y , Sun B , Yang T, Lu G, Wang X, Chai L, Li H, Lu J, Sun L, Zhang X, Ge B, Hu X, Zhang L, Hua Z, Zhang L, Zhang S, Liu N, Jiang B, Kong L, Xiang Z,Ren Y, Liu J,Jiang X, Bao Y, Zhang W, Li X, Li G, Liu G,Shen D, Sikora A, Zhai X, Zhu D,Zhang T, and Liu T (2024) Evaluation of OpenAI o1: Opportunities and Challenges of AGI <https://arxiv.org/pdf/2409.18486>

Kavlakoglu E, Stryker C (2024) What is artificial intelligence (AI)? IBM Think <https://shorturl.at/HqRdb>

Jonker A, McGrath A(2024) What is AI safety? IBM Think <https://shorturl.at/ZFmgE>

Mesut C , Dogan G,LuLu (2024) Adverse impacts of revealing the presence of “Artificial Intelligence (AI)” technology in product and service descriptions on purchase intentions: the mediating role of emotional trust and the moderating role of perceived risk , Journal of Hospitality Marketing & Management

<https://www.tandfonline.com/doi/full/10.1080/19368623.2024.2368040#abstract>

مجلة الحكمة The Dangers Of Algorithmic Bias In Artificial Intelligence .(2025) Timedjeghdine N للدراسات الفلسفية

Google Responsible AI Progress Report (2025) ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf

Soleimani M, Intezari A, Arrowsmith J, J. Pauleen D & Taskin N (2025): Reducing AI bias in recruitment and selection: an integrative grounded approach, The International Journal of Human Resource Management, DOI: 10.1080/09585192.2025.2480617

معجم المعاني الجامع <https://www.almaany.com/ar/dict/ar-ar/%D8%AA%D8%AD%D9%8A%D8%B2/>

oecd <https://www.oecd.org/en/topics/sub-issues/ai-principles.html>

Hang Li Tianlong Xu Chaoli Zhang Eason Chen Jing Liang Xing Fan Haoyang Li Jiliang Tang Qingsong Wen (2024) Bringing Generative AI to Adaptive Learning in Education <https://arxiv.org/pdf/2402.14601>

Hosanagar K, Ahn D (2024) Designing Human and Generative AI Collaboration, <https://arxiv.org/abs/2412.14199>

Kahn L, Probasco E, and Kinoshita R,(2024) AI Safety and Automation Bias, <https://cset.georgetown.edu/publication/ai-safety-and-automation-bias/>

NESCO. (2023). Recommendation on the Ethics of Artificial Intelligence. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000380455>

OpenAI. (2025). AI Risk Assessment and Mitigation Guidelines. OpenAI. <https://openai.com/research/ai-risk-guidelines>

Meta AI. (2025). Mitigating Cultural Bias in Large Language Models. Meta Research. <https://ai.facebook.com/research/publications/>

SDAIA (Saudi Data & AI Authority). (2024). Saudi AI Ethics Framework. Riyadh, Saudi Arabia. <https://sdaia.gov.sa/en/aiethics>

European Commission. (2024). Artificial Intelligence Act: Proposal for Regulation. <https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>

الملاحق:

شكل ١: النموذج العلمي للتطبيق المتبع في الدراسة