

نموذج

مشروع خطة أطروحة للتسجيل لدرجة الدكتوراة

(تخصص تقنية المعلومات)

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**

اسم الباحث: علي حسن الشهري

الرقم الأكاديمي: 20223006415

الكلية: تقنية المعلومات

جامعة ميدأوشن

Ali Hassan Alshehri

PhD Researcher, Information Technology Dept. ,Midocean University

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

للأساتذة 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 وأهدافه وفرضياته ومفاهيمه

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.

١. أزمة الموثوقية

تشير الأدبيات التجريبية الحديثة (Mesut C , Dogan G, LuLu., 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 آخذة 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 (Timedjeghdine 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. القصور في آليات التحكم

تعاي 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 لمراقبة الذكاء الاصطناعي في تحسين المستوى الأخلاقي للنظام؟
- كيف تسهم الأنظمة بعد استخدامها ضمن الأساليب والتقنيات المجربة؟، وكيف تعود الحياة الطبيعية مع المحافظة 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 وهياكل 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 نظرياً عند خروج الذكاء الاصطناعي عن السيطرة بسبب التحيز من خلال قائمة تحقق شاملة ومنظمة، يتم التحقق منها من خلال حالات واقعية، مرتبطة بالسلامة، خلال ٤ 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.

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. يمكن للذكاء الاصطناعي أن يتخذ قرارات متحيزة بناءً على مدخلات تؤثر على قراراته ومخرجاته ويزداد مستوى التحيز بزيادة التعليم الخاطئ للنظام
٣. هناك دور لكشف التحيز وتخفيفه والتعافي منه في تحسين نتائج أنشطة الذكاء الاصطناعي وكذلك تحسين معايير السلامة.

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 والحوافز 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 (معاذ 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، بدءًا من تشويه عملية اتخاذ القرار، وصولًا 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

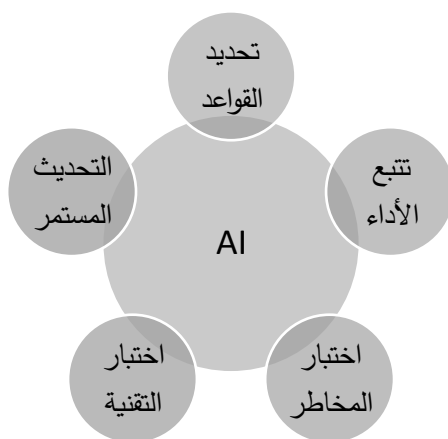
حدود المكان: بيئة الكترونية تجريبية تركز على أكثر من نظام للذكاء الاصطناعي لتجربة التحيز وتقييم الأداء.

حدود الزمان: بين عامي ٢٠٢٥ و ٢٠٢٦ م

ستقدم الدراسة إطارًا 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.

سيتم تطبيق النموذج العملي التالي:



شكل ١: النموذج العملي للتطبيق المتبع في الدراسة

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>
- معاذ، محمد (2020) كيف يحدث تحيز الذكاء الاصطناعي وما هي حلوله الممكنة؟ منصة فهم <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

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 (iia) - 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, Jiayi
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>

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

<https://asjp.cerist.dz/en/article/263149>

الملاحق:

شكل ١: النموذج العلمي للتطبيق المتبع في الدراسة