

نموذج
مشروع خطة أطروحة للتسجيل لدرجة الدكتوراة
(تخصص تقنية المعلومات)

العنوان باللغة العربية:

تطوير آلية للتعامل مع الذكاء الاصطناعي في البيئات الخارجة عن السيطرة من تقييم المخاطر إلى التخفيف والتعافي

ب. العنوان باللغة الإنجليزية:

Developing a framework for Artificial Intelligence in uncontrolled environments, 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

1446هـ - 2025 م

مع استمرار تطور تكنولوجيا الذكاء الاصطناعي والذكاء الاصطناعي العام، تزايدت المخاوف أيضًا، وتزايدت المخاوف بشأن المواقف التي تعمل فيها هذه التقنيات خارج الإشراف البشري. تقدم هذه الدراسة نهجًا للتفاعل مع الذكاء الاصطناعي في مثل هذه السياقات الحرجة تحت آليات التحكم. في المقارنة بين الذكاء الاصطناعي الضيق والذكاء الاصطناعي العام، تهدف الدراسة إلى تحديد المخاطر، وتطوير خطط الاحتواء، ووضع خطط طوارئ للتهديدات التي تشكلها عمليات الذكاء الاصطناعي غير المنضبط. من خلال مقارنة الجوانب النظرية للأبحاث، والتنبؤ بالمخاطر، وبناء السيناريوهات للتهديدات المحتملة في سياق الخدمات المقدمة بواسطة الذكاء الاصطناعي، تعمل هذه المساهمة على تعزيز مناقشة سلامة الذكاء الاصطناعي من خلال تحليل طرق تقييم المخاطر، والإجراءات التحضيرية، والجوانب الأخلاقية للتحكم في الذكاء الاصطناعي. سيساعد هذا البحث في تطوير الأساليب التنظيمية المناسبة لمختلف تقنيات الذكاء الاصطناعي، لا سيما فيما يتعلق بالحفاظ على توافقها مع القيم الإنسانية والمعايير الأمنية.

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

Abstract:

As AI and AGI 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. In comparative between Narrow AI and AGI the study aims to identify risks, develop containment plans, and develop contingency plans for threats posed by uncontrolled AI operations. By research comparison theoretical aspects, risk prediction, and scenario building for potential threats in the context of AI-provided services, this contribution advances the discussion of AI safety by analyzing risk assessment methods, preparatory procedures, and ethical aspects of AI control. This research will help develop appropriate regulatory approaches for various 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.

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). نتناول هذه الدراسة سلوك الآلة، وتركز على ثلاثة عناصر رئيسية: الكشف (لتحديد التهديدات المحتملة)؛ والوقاية (الإجراء الذي يعد حل مناسب)؛ والتعافي (لإعادة الأمور إلى نصابها). علاوة على ذلك، ستعالج هذه الدراسة أيضًا عدم موثوقية الذكاء الاصطناعي، بما في ذلك النوع الذي يعمل مثل التفكير البشري، والذي يمكنه معالجة إجراءات مثل اختراق الأنظمة أو رفض الوصول وما إلى ذلك، بالإضافة إلى تقنيات التزييف العميق المتقدمة. الذكاء الاصطناعي سلاح ذو حدين؛ إما أن يُدمر حياتنا أو يُحسّنها. علينا اتباع قواعد وإرشادات صارمة في هذا الصدد، ويجب ألا نغفل أبدًا عن الأساليب الأكثر فعالية التي نتحكم في الذكاء الاصطناعي. مشكلة الدراسة وأسئلتها

يُثير النمو السريع للذكاء الاصطناعي، وخاصةً نحو الآلات التي تُنافس ذكاء الإنسان، صعوبات حقيقية. ماذا يحدث عندما تبدأ هذه الأنظمة بفعل أشياء لم نخطط لها؟ ستجري هذه الدراسة بحثًا مباشرًا، مُستندةً إلى أفكار مُستمدة من نظرية التحكم - حيث يُمكن تعديل الأمور من خلال التغذية الراجعة للحفاظ على استقرارها، وإعدادات 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 مُبكرًا، وإيقافها بأساليب التحكم، والتحكم فيما يلي. يتحدث بعض الباحثين، مثل سواريس وآخرون (٢٠١٥)، عن سهولة إيقاف الذكاء الاصطناعي. يُشير بوستروم (٢٠١٤) إلى مخاطر كبيرة، مثل أن يُدمرنا الذكاء الاصطناعي دون استراتيجية واضحة لإيقافه.

سُطرح الدراسة الأسئلة التالية:

- ما هي المخاطر الدقيقة التي تجعل الذكاء الاصطناعي يخرج عن السيطرة؟
 - كيف نكتشف هذه المشاكل قبل أن تسيطر علينا؟
 - ما هي الأساليب والتقنيات التي توقف الذكاء الاصطناعي، وما مدى فعاليتها؟
 - 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، قد تتهاور بشدة إذا لم يتم التدخل بسرعة.
- تهدف الدراسة إلى ربط الأفكار النظرية بإصلاحات 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 إعطاء أساليب التوقف دورًا محوريًا في خطة ثلاثية الأجزاء: الكشف، والتخفيف، والتعافي، مع التركيز على نظرية التحكم في الذكاء الاصطناعي والذكاء الاصطناعي العام. كما تفتح نقاشًا حول الحوكمة الأخلاقية للذكاء الاصطناعي.

الأهمية التطبيقية:

على أرض الواقع، تقدم الدراسة لمحة فنية ودراسة حالة للباحثين ومتخصصي الذكاء الاصطناعي للتفكير في كوارث الذكاء الاصطناعي ومراقبتها وإعادة بناء الثقة، والتأكد من عدم خروج هذه الأنظمة عن السيطرة.

أهداف الدراسة:

ستضمن 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 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.

حدود الدراسة:

حدود المكان والزمان

حدود المكان: بيئة افتراضية تجريبية.

حدود الزمان: بين عامي ٢٠٢٥ و٢٠٢٦ م

ستقدم الدراسة إطاراً لدراسة سلوك الذكاء الاصطناعي ومنهجية للتعرف على الشذوذ. سيتم تفعيل استراتيجيات التخفيف، بما في ذلك آلية إيقاف أو إنهاء، للحفاظ على استقرار النظام والتحكم في الحوادث. سيتم استعادة الحالات التشغيلية باستخدام بروتوكولات الاسترداد، مع الحفاظ على سلامة النظام. يتيح اعتماد هذا الإطار المنهجي فهم ومعالجة أداء الذكاء الاصطناعي المتطور ضمن بيئة بحثية.

ستستكشف الدراسة أساليب وتقنيات للتحكم في الذكاء الاصطناعي الخارج عن السيطرة. دراسة تستند إلى مبادئ تتقاطع فيها القضايا الأخلاقية والتقنية، وذلك لتغطية طبيعة وتطبيق هذا النوع من تطبيقات الذكاء الاصطناعي، لضمان بقاء النتائج تحت السيطرة. وتشمل أهداف البحث الرئيسية ما يلي:

- تحديد نقاط الفشل وتقييم المخاطر
- 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.
- تطوير آلية فعالة للإنهاء.
- استكشاف وتقييم آليات إيقاف التشغيل التي يمكنها إيقاف عمل أنظمة الذكاء الاصطناعي عند الحاجة.
- تسليط الضوء على الحلول (مجموعات الأجهزة والبرمجيات).
- تقييم موثوقية وكفاءة وأمان طرق الإنهاء المختلفة.
- كيف ينبغي للمؤسسات التعامل مع احتواء الذكاء الاصطناعي؟

• يجب أن تكون آليات إيقاف تشغيل الذكاء الاصطناعي متوافقة مع مبادئ الذكاء الاصطناعي الأخلاقية، بحيث تُسبب أقل قدر من الضرر للمستخدمين والبنية التحتية. • البحث في مخاطر تفعيل الحلول على الأنظمة الحيوية المعتمدة على 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

• إنشاء آليات أمان متعددة الطبقات لضمان عدم تجاوز الذكاء الاصطناعي أو تعطيله.

• تطبيق بروتوكولات التكرار لضمان عدم تسبب إنهاء النظام في أعطال كارثية غير مقصودة.

• استكشاف دور الرقابة البشرية وأنظمة المراقبة الآلية في استراتيجيات احتواء الذكاء الاصطناعي.

• المساهمة في أبحاث وتطوير سياسات سلامة الذكاء الاصطناعي

• ربط المناقشات النظرية حول مخاطر الذكاء الاصطناعي بإجراءات التخفيف العملية.

• نشر مناهج علمية لاحتواء المخاطر في مجال الذكاء الاصطناعي، والتي يمكن أن تُثري سياسات حوكمة الذكاء الاصطناعي.

منهجية الدراسة:

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

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

• المنظور الأخلاقي: الأبعاد الأخلاقية والقانونية لإيقاف أنظمة الذكاء الاصطناعي عند تعطلها

منهج دراسة الحالة:

• تكوين بيئة افتراضية: كيفية عمل أنظمة الذكاء الاصطناعي لتقييم سلوكها.

• اختبار سيناريوهات الخروج عن السيطرة: إنشاء سيناريوهات يتم فيها تجاوز الحدود التشغيلية، والتحليل

المطلوب لسلوك الذكاء الاصطناعي عند استخدام آليات التوقف.

• تقييم فعالية آلية التحكم: تقييم نجاح آلية التحكم في تعطيل الأنظمة الخارجة عن السيطرة، مثل الاستجابة، وقابلية الاستغلال، وتأثير التوقف على استمرارية النظام.

دراسات مرجعية:

١. 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

ملخص البحث:

يقترح هذا البحث 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. ووفقاً لشعرات ASI، تُصنف الأجهزة وفقاً 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 نوعاً من التنسيق العالمي للإغلاق السريع.

٢. لينيك د.، ٢٠٢٠ نمط تصميم مفتاح الإيقاف لهياكل الخدمات المصغرة على أجهزة إنترنت الأشياء،

<https://dam-oclc.bac-lac.gc.ca/eng/2e811c2d-4d14-44c9-b345-44e8be2cabda>

ملخص البحث:

تقدم هذه الأطروحة نمط تصميم "مفتاح الإيقاف" لهياكل الخدمات المصغرة على أجهزة إنترنت الأشياء لتحسين الأمان وسهولة الصيانة. يكمن الجانب الجديد في هذا الحل في أنه باستخدام هذه الطريقة، يمكن لأجهزة إنترنت الأشياء غير الآمنة أن تقلل وظائف التطبيق ديناميكياً إلى الوظائف الأساسية فقط.

يُوجد هذا العمل نمط المفهوم لتطبيقات الخدمات المصغرة القائمة على المحاكاة الافتراضية، ويقترح خوارزمية لنقل وضع مستوى التشغيل، ويُثبت صحتها من خلال ثلاثة تطبيقات مختلفة: تطبيق MVC عام قائم على بنية الخدمات المصغرة، ونظام كامل يستخدم Suricata للتعريف، ومحرك Docker مُعدّل. يكمن سر هذا النمط في تحديد مستويات التشغيل لمكونات الخدمات المصغرة ومسارات اتصالها، مما يسمح لنا بإضافة أو تغيير وظائف التطبيقات ديناميكياً. وهذا يُقلل من احتمالية التعرض للهجوم، ويحمي الوظائف الحيوية من التهديدات الأمنية أو قيود الموارد. تصف الرسالة أيضاً حالات الاستخدام والسيناريوهات وعمليات تصميم التطبيقات، وتُشير إلى سبل البحث المستقبلي.

٣. ليون ب، ٢٠١٩، معالجة الشكل النهائي للتحكم في الأمن السيبراني، دراسة حالة متعددة، دراسة حالة لـ "في Y لـ"مفتاح إيقاف الإنترنت،

<https://surface.syr.edu/cgi/viewcontent.cgi?article=2015&context=etd>

تبحث هذه الرسالة في "مفتاح إيقاف الإنترنت"، أي جهود الحكومات لمقاطعة جميع أنشطة الإنترنت داخل حدودها، في الحكومات الديمقراطية والهجينة.

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 المُمكنة بين الأنظمة الديمقراطية والأنظمة الهجينة. ويتساءل عن معادلة عمليات قطع الإنترنت مع الرقابة، ويدرس القراءات المختلفة للأمن القومي. ويُقدم مساهمة جديدة في النقاش الأكاديمي من خلال دراسة دوافع وقدرات الحكومات على تبني ضوابط الأمن السيبراني.

٤. 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.

٥. هندريكس د. وأون، ٢٠٢٣، نظرة عامة على مخاطر الذكاء الاصطناعي الكارثية،

<https://arxiv.org/pdf/2306.12001>

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.

٦. تيمرز 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 تيمرز، 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 المؤلفون ثلاث استراتيجيات لمواجهة هذه التحديات: إدارة التنوع، والشراكات الاستراتيجية، والتوجه العالمي.

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 بالتفصيل، حيث يمكن رصد هذه الحوادث في قطاعات حساسة، مثل الطاقة والرعاية الصحية والنقل، إلخ. بعد ذلك، يتم تحليل الدراسة، وتحديدًا الدراسات التي أنشأت بيئة افتراضية 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، وتقييم إجراءات التخفيف مثل الطوارئ. يتم استكشاف الأساليب والتقنيات المتعلقة بالثغرات الأمنية، ووضع بروتوكولات الحوكمة.

مع حالات افتراضية لتقييم أداء النظام، وتسجيل معلومات الاستقرار. كما تتم مناقشة الجوانب الأخلاقية والقانونية لآليات الإغلاق بشكل أعمق مع مراعاة الأعمال السابقة ذات الصلة، ويتم استخلاص توصيات لهذا الغرض للجهات التنظيمية والمطورين على حد سواء. ثم يتم استخدام تقنيات إحصائية لتحليل البيانات وتقديم اقتراحات عملية لتعزيز سلامة الذكاء الاصطناعي.

دراسة حالات محددة وتحليل البيانات.

- تحليل قائم على الحالات لفشل أنظمة الذكاء الاصطناعي في قطاعات مثل الكهرباء والرعاية الصحية والنقل والمركبات ذاتية القيادة.
- تحليل سلوكي وتحليل النظام للتحقق من قدرات النظام.
- تصميم توصيات ملموسة لوضع لوائح فعالة بناءً على نتائج التحليل.

أساليب جمع بيانات الدراسة:

تعتمد هذه الدراسة نهجًا مقارنًا بين التعامل مع الذكاء الاصطناعي الضيق والعام في حالات فقدان السيطرة. 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، مع الأخذ في الاعتبار أساليب التحكم المختلفة.

مراجعة الدراسات السابقة:

- دراسة شاملة للدراسات السابقة حول آليات التحكم بالذكاء الاصطناعي 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.
- آليات التحكم بالذكاء الاصطناعي المطبقة بنجاح في الدراسات التجارية ودراسات الحالة.

• تحديث التشريعات واللوائح.

• مراجعة لقوانين ولوائح سلامة وضبط الذكاء الاصطناعي، ومقارنة بين مختلف الحالات.

• يُقيم هذا البحث آثار هذه الأطر القانونية على تطوير ونشر الذكاء الاصطناعي.

التحليل المقارن:

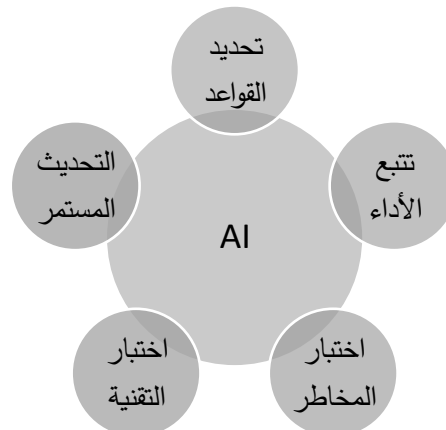
- تحليل بيانات دراسات الحالة والأدبيات، وتطوير إطار عمل مقارن.
- تقدم الدراسة إطار عمل مقارن لتقييم مزايا وعيوب آليات ضبط الذكاء الاصطناعي الحالية.
- ستساعد الدراسة المستخدمين والمطورين، وتتصّحهم بتصميم أنظمة الذكاء الاصطناعي الخاصة بهم بشكل آمن ومرن لمواجهة هذه الثغرات.

النتائج المتوقعة:

- صياغة منهجية شاملة لتقييم المخاطر وسيناريوهات الخلل المرتبطة بالذكاء الاصطناعي.
- التوصية بتحسينات في سياسات أمن الذكاء الاصطناعي والتحكم فيه.
- تقديم أدلة تجريبية لمساعدة صانعي السياسات والمطورين على بناء أنظمة ذكاء اصطناعي أكثر أمانًا.
- ستساعد الدراسة المستخدمين والمطورين، وستتصّحهم بتصميم أنظمة الذكاء الاصطناعي الخاصة بهم بشكل آمن ومرن لمواجهة هذه الثغرات.
- فهم السمات والعوامل والنتائج الشائعة لحالات فقدان التحكم في الذكاء الاصطناعي.

كيف سيتم تنفيذ الجانب العملي من الدراسة:

سيتم تطبيق النموذج العملي التالي:



استخدام أدوات محاكاة لسلوك الذكاء الاصطناعي ومراقبته كنقطة انطلاق:

TensorFlow: مكتبة مفتوحة لتطوير نماذج ذكاء اصطناعي واختبار سلوكها في بيئات محاكاة.

PyTorch: أداة مشابهة لـ TensorFlow، تتيح إنشاء نماذج ذكاء اصطناعي بسيطة ومراقبة قراراتها.

OpenAI Gym: بيئة لمحاكاة أنظمة ذكاء اصطناعي واختبار تصرفاتها في سيناريوهات مختلفة

قائمة المراجع:

References:

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv preprint arXiv:1606.06565.

Bekkum M , Borgesius F (2022) Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? <https://www.sciencedirect.com/science/article/pii/S0267364922001133>

Bekkum M, Borgesius F, 2023 Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? <https://rb.gy/e9ivs2>

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 Fitzgerald, Aaron Boddy, & Seth D. Baum 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. (2015). Corrigibility. Workshops at the 29th AAAI Conference on Artificial Intelligence.

<https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124>

Floridi, L., & Cowls, 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.

Yampolskiy, R. V. (2016). Artificial Intelligence Safety and Security. CRC Press.

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

H. F. Atlam and G. B. Wills, (2020) Iot security, privacy, safety and ethics,” Digital twin technologies and smart cities, pp. 123–149,.

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

M (2024) Enhancing Cybersecurity With Artificial Immune Systems and General Intelligence: A New Frontier in Threat Detection and Response. IEEE Access.

F. Mo, Z. Tarkhani, and H. Haddadi, “Sok: (2022)machine learning with confidential computing,” arXiv preprint arXiv:2208.10134,.

Lin Zhao et al. (2023).When brain-inspired ai meets agi. In: arXiv preprint arXiv:2303.15935

Fei Dou et al. (2023) Towards Artificial General Intelligence (AGI) in the Internet of Things (IoT):Opportunities and Challenges”. In: arXiv preprint arXiv:2309.07438

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 Holmes , YanZhuang , Jiaqi Wang , 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, HuaZhou, 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>