الحوكمة وإدارة المخاطر في أنظمة الذكاء الاصطناعي وفق المعايير الدولية

ورقة مقدمة من الاتحاد العربي للاقتصاد الرقمي
عضو الفريق العربي للذكاء الاصطناعي

د. أيمن غنيم
الأمين العام المساعد

أهمية تطبيق معايير الذكاء الاصطناعي

- التطور السريع للذكاء الاصطناعي يتطلب ضوابط أخلاقية وتنظيمية.

- المخاطر تشمل المسؤولية، وضوح سبب القرارات، التحيز، الخصوصية، والأمان السيبراني.

- المشرعون والحكومات عليهم وضع الأطر التنظيمية المناسبة لضمان الاستخدام المسؤول.

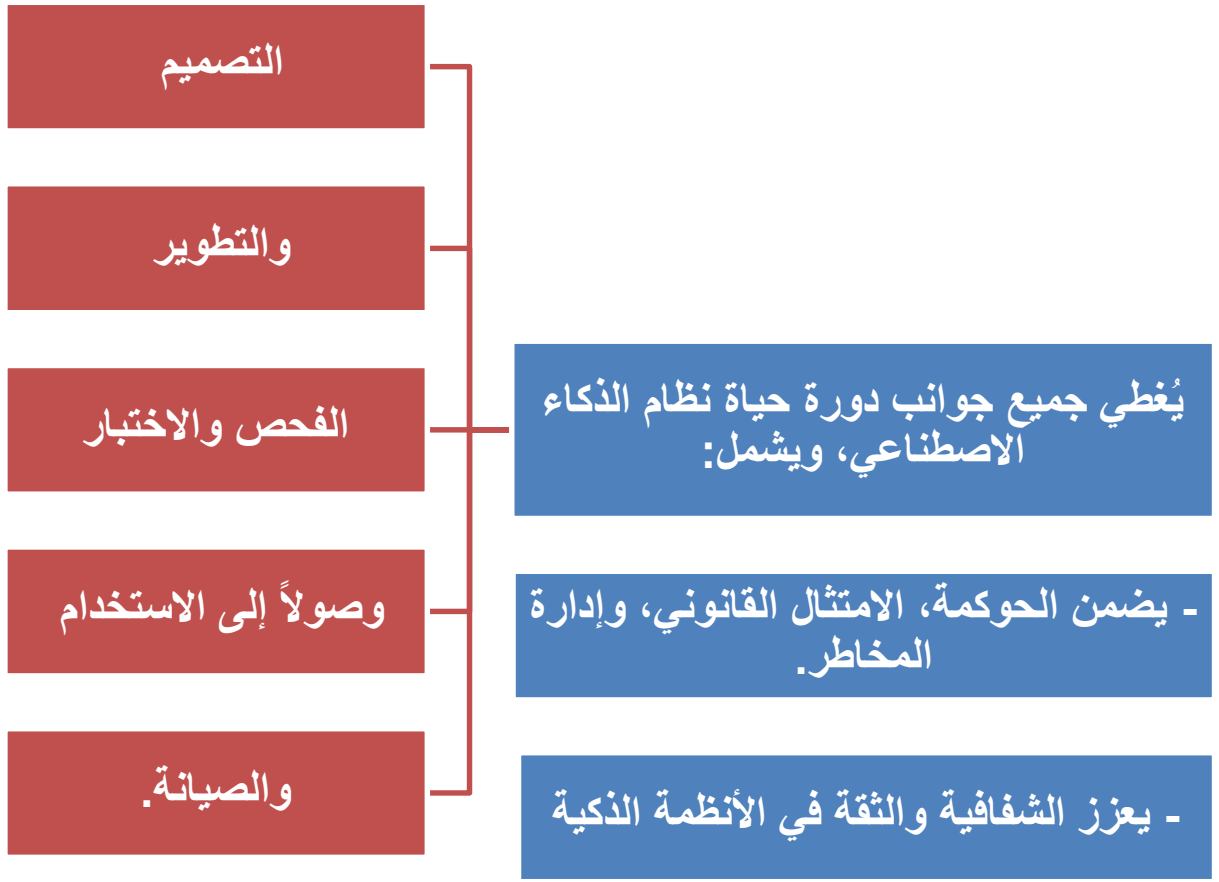
ما هو دور صناع القرار/المشرعين ... في ضبط الذكاء الاصطناعي

تبني قوانين وتشريعات منظمة تسمح بتشجيع الابتكار وبنفس الوقت حماية المجتمع من أي استخدامات ضارة

إيجاد وسيلة قابلة للتطبيق والتقييم والمحاسبة مثل المواصفات والمعايير لإنفاذ تلك القوانين والسياسات.

- مراقبة تطبيق هذه المعايير في مختلف مراحل دورة حياة الذكاء الاصطناعي من التطوير وانتهاء بالتطبيقات

نظام إدارة الذكاء الاصطناعي ISO/IEC 42001



إدارة مخاطر الذكاء الاصطناعي ISO/IEC 23894 -

إطار عمل تطوير الأنظمة الذكية باستخدام التعلم الآلي - ISO/IEC 23053

- يؤكد على المطور لتضمين شفافية طريقة عمل الأنظمة ، لتكون أكثر تفسيرية وشفافة

- يركز على ضمان جودة النماذج وموثوقيتها قبل نشرها.

- يوفر إطارًا شاملًا لوصف وتصميم وتطوير أنظمة الذكاء الاصطناعي.

هذه المواصفات أجمعت على تلك المتطلبات التكنولوجية والبشرية

خوارزميات معتمدة: عند الحاجة لاستخدام مصادر مفتوحة فيجب مراعاة موثوقيتها.

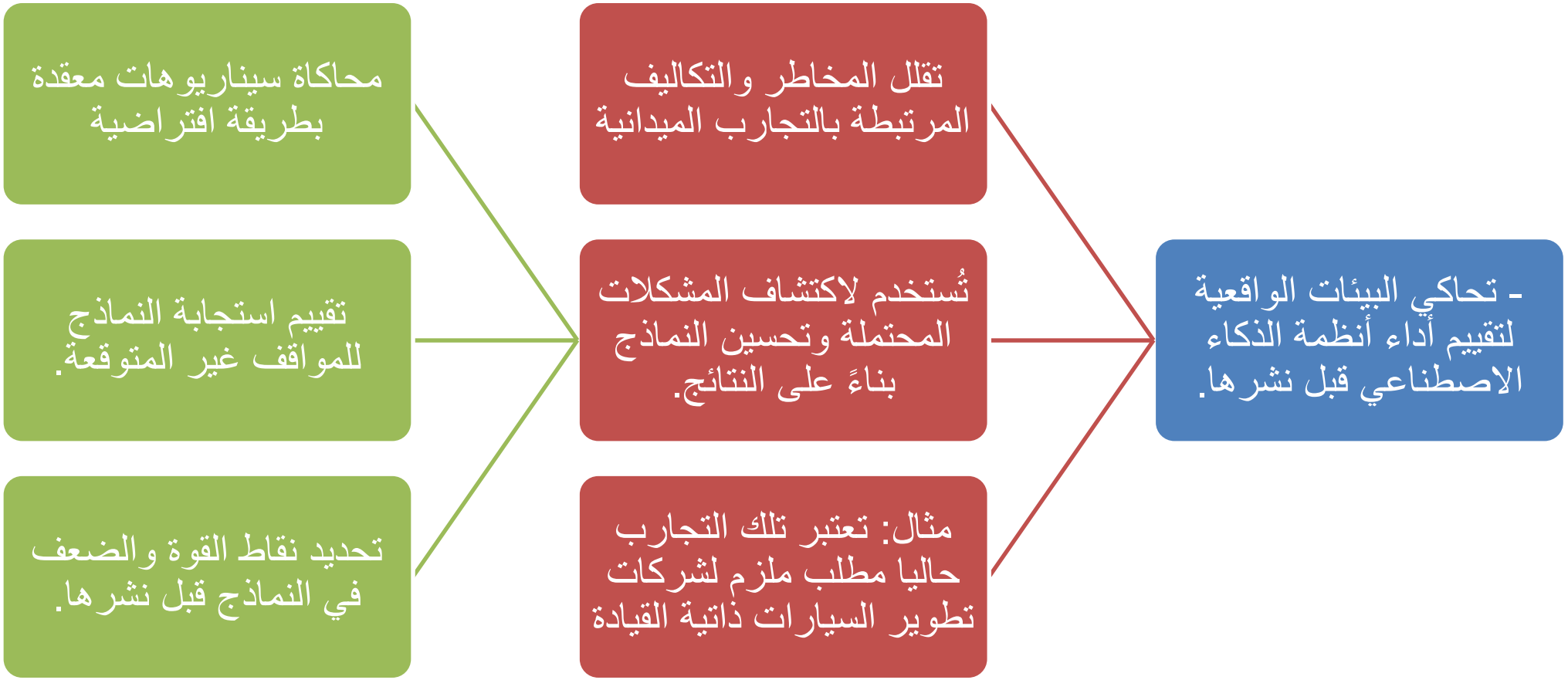
تأهيل المطورين: ضمان كفاءة المطورين في تصميم وتقييم الأنظمة.

الشفافية والمساءلة: توثيق العمليات واتخاذ القرارات بوضوح.

إدارة المخاطر: تحديد المخاطر المحتملة وتأمين الأنظمة ضد الهجمات.

بيئة اختبار متقدمة: توفير أدوات متطورة لضمان جودة الأداء.

اختبارات المحاكاة في الذكاء الاصطناعي virtual simulation



اختبارات التطبيق الميداني في الذكاء الاصطناعي

قبل إطلاق التطبيق او المنتج المذي يستخدم الشكاء الاصطناعي، يجب مروره باختبارات التطبيق الميداني.

من المهم تجربة عمل تلك الأنظمة في بيئات حقيقية لقياس الأداء الفعلي.

- اختبارات الإجهاد: تقييم استجابة الأنظمة تحت الضغوط العالية.
- اختبارات التحيز: كشف وتصحيح التحيزات في البيانات والنماذج.
- اختبارات الأمان: ضمان مقاومة الأنظمة للهجمات السيبرانية.

علاقة المواصفات بالمؤسسات المستفيدة والمطورة

على المؤسسات المستفيدة الامتثال لحوكمة الإجراءات الإدارية ذات العلاقة، حماية حقوق المستخدمين، وإدارة المخاطر.

على المؤسسات المطورة ضمان جودة النماذج، اختبارها، توفير الشفافية، وتضمين المنظومة إصدار تفسير واضح لقراراتها.

أدوات عالمية لدعم حوكمة الذكاء الاصطناعي

الأداة	الجهة المطورة	أهم الخصائص / ما تختبره	التدقيق على الكود؟	ارتباطها بمتطلبات الحوكمة
AI Fairness 360	IBM Research	- تكتشف التحيز في البيانات والنماذج عبر 70+ مقياسًا. - توفر حلولًا لتخفيف التحيز (مثل إعادة وزن البيانات).	تقيّم النتائج فقط	ISO 23894) إدارة المخاطر) - معالجة التحيز وضمان الإنصاف.
Amazon SageMaker Clarify	AWS (Amazon)	- تحلل الانحياز في البيانات وتفسير قرارات النماذج. - توليد تقارير شفافية تلقائية.	تقيّم النتائج فقط	ISO 42001) الحوكمة) - توثيق الشفافية والامتثال.
TensorFlow Responsible AI	Google	- أدوات تفسيرية مثل **What-If Tool** لفحص سلوك النماذج. - تحليل تأثير التغييرات في البيانات على النتائج.	تقيّم النتائج فقط	ISO 23053) التطوير) - ضمان تفسيرية النماذج وجودتها.
Fairlearn	Microsoft	- تقييم الإنصاف في النماذج عبر مقاييس مثل **Demographic Parity**. -اقتراح تعديلات لتحقيق العدالة.	تقيّم النتائج فقط	ISO 23894 - معالجة مخاطر التحيز في الأنظمة.
IBM Watson OpenScale	IBM	- مراقبة الأداء واكتشاف الانحرافات في النماذج المُنشأة. - شرح القرارات المعقدة بلغة بسيطة.	تقيّم النتائج فقط	ISO 42001 - مراقبة الامتثال التشغيلي والشفافية.
H2O Driverless AI	H2O.ai	- تطوير نماذج قابلة للتفسير تلقائيًا. - توثيق خطوات بناء النموذج (.Feature Engineering)	تقيّم النتائج فقط	ISO 23053 - ضمان جودة النماذج وقابليتها للتدقيق.
Themis-ML	أكاديمية / مفتوحة المصدر	- اختبار الإنصاف في نماذج التصنيف (مثل: القروض، التوظيف). - دعم معايير العدالة الدولية.	تقيّم النتائج فقط	ISO 23894 - الامتثال لمعايير العدالة الدولية.
LIME (Local Interpretable Model-agnostic Explanations)	جامعة واشنطن (مفتوحة المصدر)	- تفسير قرارات النماذج المعقدة (مثل: لماذا رُفض طلب قرض؟). - توضيح تأثير العوامل الرئيسية على القرار.	تقيّم النتائج فقط	ISO 23053 - تعزيز تفسيرية النماذج.
Deepchecks	Deepchecks	- فحص سلامة البيانات والنماذج (مثل: تسرب البيانات، توزيع غير متوازن). - مراقبة الانحرافات أثناء التشغيل.	تفحص الكود والنتائج (محدودًا)	ISO 42001 - ضمان جودة البيانات والامتثال التشغيلي.
CodeCarbon	Mila، BCG، Haverford College	- قياس البصمة الكربونية لتدريب وتشغيل نماذج الذكاء الاصطناعي. - اقتراح تحسينات للكفاءة البيئية.	تفحص الكود (حساب استهلاك الموارد)	ISO 42001 - الامتثال لمعايير الاستدامة والمسؤولية البيئية.
SHAP (SHapley Additive exPlanations)	جامعة واشنطن (مفتوحة المصدر)	- تفسير قرارات النماذج عبر قيم "شابلي" الرياضية. - تحديد مدى مساهمة كل متغير في القرار النهائي.	تقيّم النتائج فقط	ISO 23053 - تعزيز الشفافية في اتخاذ القرارات.

تأسيس مركز عربي لتدقيق أنظمة الذكاء الاصطناعي :التوصية

الهدف	الدور الرئيسي
مراقبة الامتثال	فحص الأنظمة باستخدام أدوات دولية (مثل AI Fairness 360) للتأكد من تطبيق معايير ISO.
إصدار شهادات اعتماد	منح شهادات "معتمد عربيًا" للأنظمة الملتزمة بالمعايير الأخلاقية والأمان.
تطوير أدوات محلية	تصميم منصات تدقيق بلغة عربية لدعم الشركات الناشئة والصغيرة.
بناء الكوادر العربية	تدريب مدققين متخصصين في مراجعة أنظمة الذكاء الاصطناعي عبر برامج معتمدة.
جذب الاستثمارات	زيادة ثقة الشركات العالمية بالسوق العربية عبر شهادات معتمدة.

شكرا لحسن استماعكم

- مع تحيات الاتحاد العربي للاتحاد الرقمي
- عضو الفريق العربي للذكاء الاصطناعي