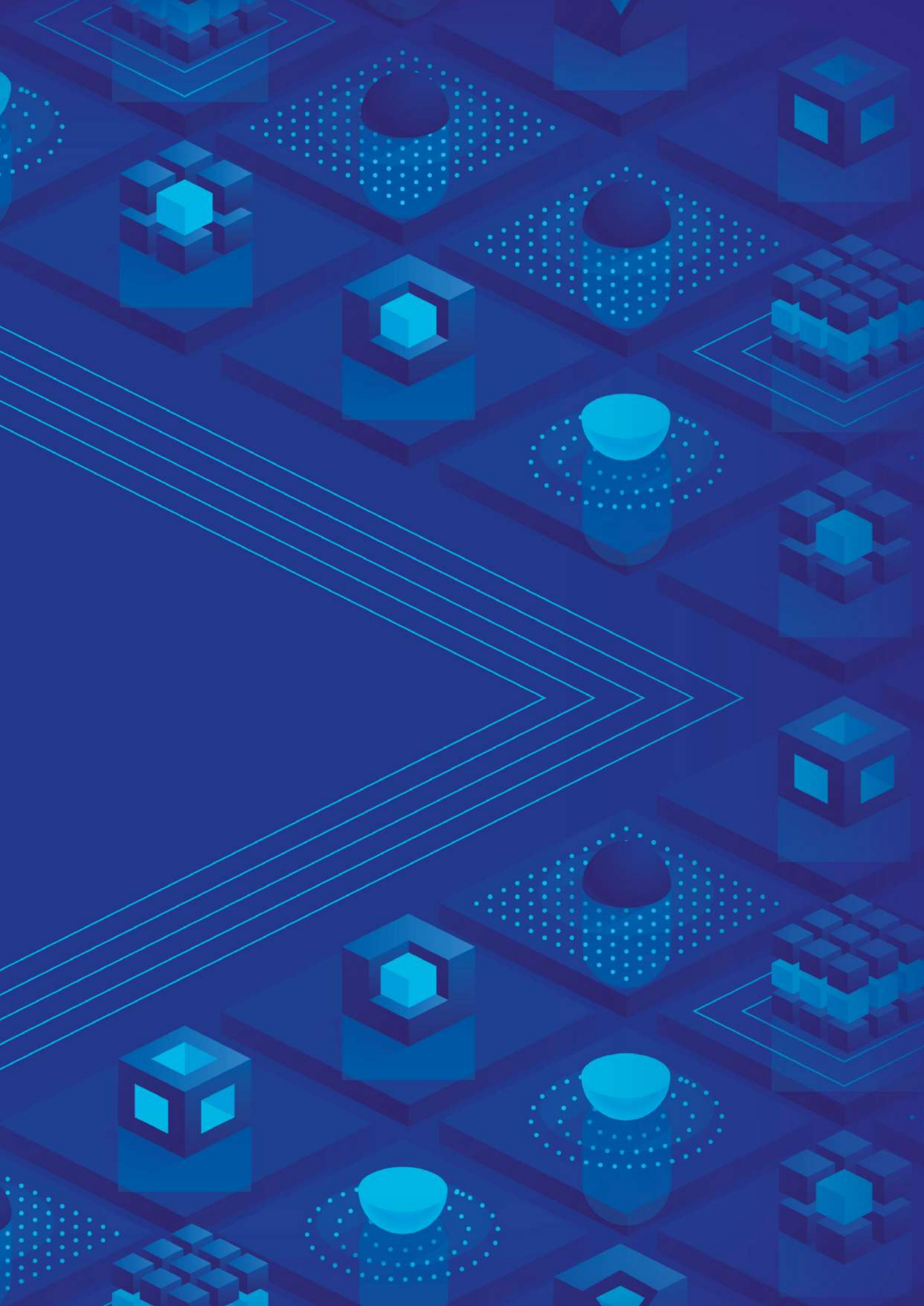
البيانات المصطنعة

Synthetic Data

الفرص والتحديات وآليات الحوكمة

منقول
عبر
afaqb.co

سبتمبر 2026م



الملخص التنفيذي

تتناول هذه الدراسة البيانات المصطنعة (Synthetic Data) باعتبارها أداة مهمة لتمكين الابتكار وحماية الخصوصية، لكنها في الوقت ذاته تمثل مصدراً لمخاطر جديدة تستلزم تطوير سياسات دقيقة ومتوازنة. وتناقش الدراسة حوكمة البيانات المصطنعة مرتكزة على أربعة أبعاد رئيسية وهي: حماية الخصوصية، وجودة البيانات، والعدالة وتخفيف التحيز، والشفافية. كما تستعرض المشهد التنظيمي الدولي لحوكمة البيانات المصطنعة، محللةً الأطر القانونية، والإرشادات التقنية، وآليات الحوكمة المعمول بها عالمياً.

يظهر التحليل أن المشهد الدولي لا يزال يفتقر إلى إطار قانوني صريح ومتكامل ينظّم البيانات المصطنعة، فأغلب الأنظمة القانونية ما زالت تتعامل مع البيانات المصطنعة ضمن الأطر القائمة لحماية البيانات الشخصية الحقيقية وتنظيم الذكاء الاصطناعي، ويُحدَّد وضعها القانوني في الممارسة العملية غالباً بناء على مستوى خطر إعادة التعرف على هوية الأفراد منها؛ فكلما استُبعدت إمكانية ربط هذه البيانات بشخص محدد، عوملت على أنها بيانات مجهولة الهوية، وإلا خضعت لقوانين حماية البيانات الشخصية. كما تُظهر التجارب الدولية التوجه نحو الأطر الإرشادية الطوعية والأدلة التشغيلية كأدوات مرنة لتشجيع التبنّي الآمن والتقليل من مخاطر هذه البيانات، إلى جانب البيئات التنظيمية التجريبية (Regulatory Sandboxes) المخصصة للبيانات المصطنعة كإحدى آليات اختبار هذه البيانات ضمن بيئات خاضعة للرقابة. وتوصي هذه الإرشادات في مجملها بأهمية إجراء اختبارات لمخاطر إعادة التعرف، وتشجيع استخدام تقنيات تعزيز الخصوصية مثل الخصوصية التباينية، إضافة إلى أهمية تطبيق متطلبات الشفافية مثل تسمية المحتوى المصطنع وتتبع مصدره، وتوثيق عمليات التوليد، والإفصاح عند استخدام البيانات المصطنعة في تدريب أنظمة الذكاء الاصطناعي.

ورغم التقدم الملحوظ في المجال، تكشف الدراسة عن فجوات تنظيمية جوهرية، تشمل غياب تعريف موحّد للبيانات المصطنعة، وعدم وضوح وضعها القانوني، وافتقار الأطر القائمة إلى معايير كمية لقياس الخصوصية والجودة والعدالة، إضافةً إلى الغموض في تحديد المسؤولية القانونية في حال وقوع ضرر نتيجة استخدام بيانات مصطنعة غير دقيقة أو متحيزة. ختاماً، تقدم الدراسة خارطة طريق عملية لتبني البيانات المصطنعة بشكل مسؤول، ترتكز على ثلاث مراحل مترابطة توازن بين الابتكار وحماية الخصوصية. تبدأ بمرحلة التأسيس الاستراتيجي التي تتضمن تحديد أهداف الاستخدام بوضوح وإنشاء إطار حوكمة داخلي يضمن المساءلة وإدارة المخاطر، تليها مرحلة التطوير التقني التي تركز على وضع معايير للتحقق من جودة البيانات وتطبيق آليات شفافية مثل العلامات المائية والبيانات الوصفية لتتبع مصدر البيانات وضمان الامتثال، ثم مرحلة بناء المنظومة من خلال بيئات اختبار تنظيمية تدعم الابتكار وتسمح بتجربة الحلول في إطار آمن، إلى جانب تطوير سياسات مرنة تُراجع دورياً لمواكبة التطورات التقنية. تؤكد الخارطة على أن البيانات المصطنعة تمثل فرصة استراتيجية لتعزيز الابتكار وحماية الخصوصية في آن واحد، شريطة دمج القيم الأخلاقية والحوكمة في جميع مراحل دورة حياتها لضمان استخدامها الآمن والمستدام.

المحتويات

المصطلحات	**7**
المقدمة	**9**
أولاً: نظرة عامة	**11**
تاريخ البيانات المصطنعة	11
مفهوم البيانات المصطنعة	11
أنواع البيانات المصطنعة	14
آلية إنشاء البيانات المصطنعة	15
تطبيقات البيانات المصطنعة	18
ثانياً: الفرص والمخاطر	**19**
الفرص والإمكانات	19
المخاطر والتحديات	20
ثالثاً: المشهد العالمي	**21**
حوكمة البيانات المصطنعة	21
الأبعاد التنظيمية	21
التجارب الدولية	24
أبرز الاتجاهات	34
الفجوات التنظيمية	35
رابعاً: الوضع الراهن في المملكة	**37**
جهود المملكة في بناء منظومة البيانات الوطنية	37
الإطار التنظيمي وعلاقته بالبيانات المصطنعة	37
خامساً: الدروس المستفادة وخارطة الطريق	**39**
الدروس المستفادة	39
خارطة الطريق	40
التوصيات	**42**
الخاتمة	**43**
المراجع	**44**

المصطلحات

المصطلح بالعربية	المصطلح بالإنجليزية	التعريف الموجز
إخفاء البيانات	Data Masking	عملية تعتيم المعلومات الحساسة عن طريق إخفاء أو تشفير القيم الحساسة (كالأسماء والمعرّفات) مع بقاء السجلات الأصلية قائمة.
إخفاء بيانات الهوية	Data Anonymization	عملية تعديل معلومات التعريف الشخصية أو إزالتها من مجموعات البيانات؛ لحماية خصوصية الأفراد.
اختبار الفصل	Singling Out	أحد اختبارات إعادة التعرف؛ يقيس القدرة على تمييز سجل واحد يخصّ فرداً محدداً ضمن مجموعة البيانات.
اختبار كشف التحيز	Bias Detection	إجراءات تحليلية لمقارنة الخصائص التوزيعية للبيانات المصطنعة بالبيانات الحقيقية للكشف عن انحرافات أو تحيّزات غير مقصودة قد تنتجها نماذج التوليد.
اختبار المتسلل المحفَّز	Motivated Intruder Test	منهجية تقييم مخاطر إعادة التعرف؛ تفترض وجود خصم يمتلك الدافع والوسائل الكافية لمحاولة تحديد هوية الأفراد من البيانات.
الاستنتاج	Inference (Attribute Inference)	استنتاج معلومات جديدة عن فرد من أحد اختبارات إعادة التعرف؛ يقيس إمكانية بيانات تبدو مجهولة الهوية.
آلية الانتباه الذاتي	Self-Attention	طبقة في الشبكة العصبية تُخرج تسلسلاً جديداً من التضمينات، عن طريق تطبيق آلية الانتباه على المعلومات الواردة من عناصر تسلسل المدخلات.
البيانات فائقة الاصطناع	Hypersynthetic Data	اتجاه مستقبلي يدمج المحاكاة الفيزيائية، والتوائم الرقمية، والنماذج التوليدية لاستكشاف سيناريوهات مستقبلية معقدة.
بيانات السلاسل الزمنية	Time-Series Data	هي نوع من البيانات يرتبط بعامل الزمن وتسلسل القيم، يسمح بتعليم الخوارزميات الأنماط الزمنية، والتنبؤ بالقيم المستقبلية، واكتشاف الشذوذ والانحرافات، ويُستخدم في تطبيقات مثل التنبؤات المالية، ومراقبة المعدات والأنظمة الصناعية، والروبوتات وإنترنت الأشياء.
البيانات المصطنعة	Synthetic Data	هي البيانات المُولَّدة بهدف محاكاة الخصائص الإحصائية والأنماط والعلاقات الكامنة في البيانات الحقيقية، وتُولد هذه البيانات بعدة أساليب، بما في ذلك النماذج الإحصائية التقليدية، وتقنيات الذكاء الاصطناعي التوليدي وأساليب المحاكاة، وغيرها من تقنيات التوليد المناسبة لطبيعة البيانات وحالة الاستخدام.
بيانات مصطنعة بالكامل	Fully Synthetic Data	تُنشأ كلياً بالذكاء الاصطناعي أو النماذج الإحصائية أو المحاكاة، دون أي بيانات حقيقية ضمنها سجلات.
بيانات مصطنعة جزئياً	Partially Synthetic Data	كأساس، ثم تُستبدل خصائص محددة (غالباً بيانات تستخدم مجموعة بيانات حقيقية حساسة) بقيم مولَّدة اصطناعياً.
بيانات مصطنعة هجينة	Hybrid Synthetic Data	مزيج من سجلات مصطنعة بالكامل وسجلات حقيقية ضمن مجموعة البيانات نفسها.
تدهور النموذج	Model Collapse	تدهور جودة البيانات التي تولدها نماذج الذكاء الاصطناعي مع مرور الوقت؛ بسبب تدريب النماذج على بيانات أنتجتها نماذج ذكاء اصطناعي.

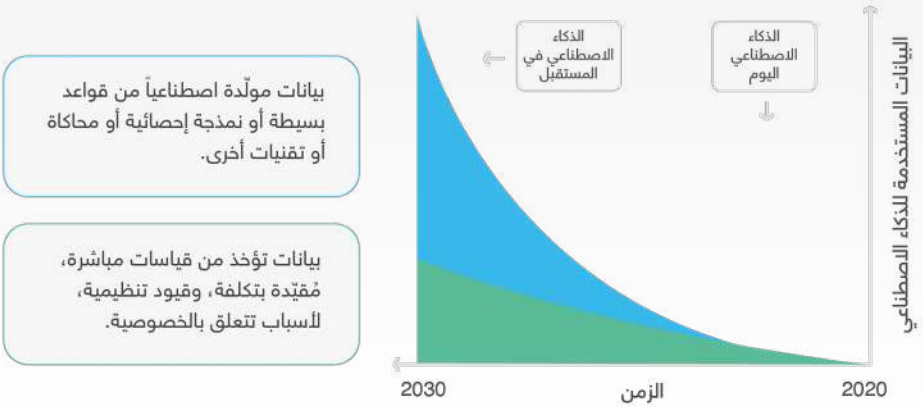
المصطلح بالعربية	المصطلح بالإنجليزية	التعريف الموجز
التطابق العرضي	Coincidental Matching	حالة تتطابق فيها بيانات مصطنعة مع بيانات فرد حقيقي بمحض الصدفة الإحصائية، دون أن يكون ذلك ناتجاً عن حفظ النموذج لأنماط فعلية.
تقنيات تعزيز الخصوصية	Privacy-Enhancing Technologies (PETs)	مجموعة تقنيات (منها البيانات المصطنعة والخصوصية التباينية) تهدف إلى تمكين الاستفادة من البيانات مع تقليل مخاطر الكشف عن هوية الأفراد.
خصوصية تباينية	Differential Privacy	تعريف رياضي للخصوصية في تعلم الآلة يتيح جمع الإحصاءات حول مجموعة البيانات وتحليلها ومشاركتها بناءً على البيانات الشخصية مع حماية خصوصية الأفراد.
خطر إعادة التعرف	Re-identification Risk	احتمال ربط بيانات مصطنعة أو مجهولة الهوية بأفراد حقيقيين ومحددين، خصوصاً حين تحفظ نماذج التوليد أنماطاً دقيقة أو نادرة من البيانات الحقيقية.
الربط	Linkability	أحد اختبارات إعادة التعرف؛ يقيس إمكانية ربط سجلين أو أكثر بالشخص نفسه عبر مجموعات بيانات مختلفة.
مثلث الجودة	Quality Triangle	إطار يوازن بين ثلاثة عناصر أساسية: المنفعة (Utility)، والمصداقية(Fidelity) ، والخصوصية (Privacy).
المصداقية	Fidelity	مدى احتفاظ البيانات المصطنعة بالتوزيعات والعلاقات الإحصائية الموجودة في البيانات الحقيقية.
المنفعة	Utility	القيمة التحليلية والفائدة الوظيفية للبيانات المصطنعة ومدى نجاحها عند استخدامها في مهام واقعية.
نموذج انتشاري	Diffusion Models	نموذج توليدي يتعلم كيفية إنشاء بيانات جديدة، عن طريق عكس عملية إضافة التشويش إلى بيانات التدريب.
نموذج متعدد الصيغ	Multimodal Model	نموذج له مدخلات أو مخرجات بصيغ متعددة.

المقدمة

شهد العالم في السنوات الأخيرة تقدماً كبيراً في قدرات الذكاء الاصطناعي، لا سيما مع الطفرة التي أحدثها تطور الذكاء الاصطناعي التوليدي (Generative AI - GenAI)، وما صاحبها من توسع في الحاجة إلى البيانات باعتبارها مكوناً أساسياً لتدريب نماذج الذكاء الاصطناعي المتقدمة. غير أن هذا التوسع في الحاجة إلى البيانات يواجه عدداً من التحديات، من أبرزها ندرة البيانات عالية الجودة والملائمة لأغراض التدريب المتخصصة، وارتفاع تكاليف جمعها وتسميتها، إلى جانب صدور تشريعات متخصصة بحماية البيانات والخصوصية، وفي مقدمتها اللائحة العامة لحماية البيانات (GDPR)، والتي تضع قيوداً قانونية دقيقة على معالجة البيانات الشخصية ومشاركتها[1]، ما يحدّ من استخدام هذه البيانات في تطبيقات الذكاء الاصطناعي. وضع هذا الواقع القائمين على مبادرات التحول الرقمي أمام تحديات جوهرية، خاصة في القطاعات الحساسة مثل الرعاية الصحية والمالية، حيث تتداخل الحاجة الملحّة إلى البيانات مع متطلبات قانونية وأخلاقية معقدة. فعلى سبيل المثال، قد تحتاج مؤسسة صحية إلى تدريب نظام ذكاء اصطناعي لاكتشاف أمراض نادرة، إلا أن القيود على بيانات المرضى تجعل مشاركة البيانات الطبية اللازمة أمراً بالغ الصعوبة. وبالمثل، قد تسعى مؤسسة مالية إلى تطوير نماذج للكشف عن عمليات الاحتيال الدولية، ولكن قيود نقل البيانات بين الدول تخلق تعقيدات قانونية تحول دون ذلك.

في ظل هذه التحديات برزت البيانات المصطنعة (Synthetic Data) كحل تقني وتنظيمي يهدف إلى التوفيق بين الحاجة إلى البيانات ومتطلبات حماية الخصوصية والامتثال، تُولَّد هذه البيانات باستخدام أساليب إحصائية متقدمة أو تقنيات الذكاء الاصطناعي التوليدي، ويكمن جوهر فائدتها في قدرتها على محاكاة الخصائص الإحصائية والهيكلية للبيانات الحقيقية دون أن تحتوي على معلومات مرتبطة بأشخاص أو أحداث فعلية، وبفضل هذه الخصائص، تُمثل البيانات المصطنعة خطوة مهمة نحو تمكين الابتكار المسؤول المعتمد على البيانات. إذ أصبحت اليوم توفّر حلاً للنقص في بيانات تدريب نماذج الذكاء الاصطناعي، وتساهم في تسهيل مشاركة البيانات بين الجهات المختلفة. كما اكتسبت اهتماماً في قطاعات مثل التمويل والرعاية الصحية، حيث تكون البيانات محدودة أو محمية، وتشير التقديرات إلى أن البيانات المصطنعة ستصبح عنصراً أساسياً في منظومة الذكاء الاصطناعي خلال السنوات المقبلة، إذ يُتوقع أن تتجاوز البيانات المصطنعة البيانات الحقيقية في تدريب نماذج الذكاء الاصطناعي بحلول عام 2030م[2] كما **يوضح الشكل (1)**.

الشكل (1): تشير التقديرات إلى أنه بحلول عام 2030م ستكون غالبية البيانات المستخدمة لتدريب نماذج الذكاء الاصطناعي مُصطنعة. المصدر: جارتنر (Gartner)[2].



ورغم ما تحمله البيانات المصطنعة من فوائد استراتيجية، فإن استخدامها يثير مجموعة من الإشكاليات التي قد تختلف بطبيعتها عن تلك المرتبطة بالبيانات

التقليدية. وتشمل قضايا الخصوصية مثل خطر إعادة التعرّف (Re-identification Risk) المتمثل في القدرة على تحديد هوية شخص محدد من هذه البيانات، وإعادة إنتاج التحيزات غير العادلة في البيانات الحقيقية، إضافة إلى الضبابية القانونية للبيانات المصطنعة ومدى انطباق قوانين حماية البيانات الحقيقية عليها. لذا تبرز الحاجة الملحّة إلى أطر حوكمة واضحة تنظم دورة حياة هذه البيانات، بدءاً من توليدها، مروراً باستخدامها ومشاركتها، وصولاً إلى تقييم آثارها، خاصة مع التوسع في استخدام البيانات المصطنعة في القطاعين العام والخاص، وتشمل هذه الأطر تقييماً منهجياً للمخاطر، ومعايير دقيقة للجودة، ومتطلبات واضحة للتوثيق والشفافية، وآليات مساءلة تضمن الاستخدام المسؤول والمتسق مع القوانين ذات الصلة. ويتقاطع هذا المسار مع منظومات أوسع لحوكمة البيانات، وسياسات مشاركة البيانات، ومعايير الأمن والخصوصية.

انطلاقاً من هذا الواقع، تهدف هذه الدراسة إلى تحليل الأطر القائمة لحوكمة البيانات المصطنعة، واستعراض الممارسات الدولية الرائدة، ورصد الاتجاهات والسياسات الناشئة عالمياً. كما تسعى إلى تشخيص الفجوات التنظيمية ودعم صانعي القرار من خلال تقديم قراءة متوازنة تجمع بين الاعتبارات التقنية والقانونية، واقتراح مبادئ عملية تساعد على الاستخدام المسؤول والمستدام للبيانات المصطنعة. وترتبط الدراسة بالسياقين الوطني والعالمي على حد سواء، إدراكاً للدور المحوري الذي يمكن أن تؤديه البيانات المصطنعة في بناء منظومة بيانات حديثة وآمنة تعزز الابتكار مع الحفاظ على قيم الحوكمة، لتُسهم في تحقيق أهداف التحول الرقمي، مع الحفاظ على ثقة المجتمع وحقوق أفراده. في النهاية، تُقدّم الدراسة خارطة طريق عملية لمساعدة الجهات الحكومية والخاصة على تبنّي البيانات المصطنعة بأعلى مستويات الكفاءة والأمان.

أولاً: نظرة عامة

في ظل التحول المتسارع نحو الاعتماد على البيانات، برزت البيانات المصطنعة كأحد الحلول المبتكرة للتغلب على قيود البيانات الحقيقية. يقدم هذا الفصل نظرة عامة على البيانات المصطنعة من حيث نشأتها وتطورها، وتعريفاتها المختلفة، وأوجه الاختلاف بينها وبين البيانات الحقيقية. كما يستعرض الأنواع المختلفة للبيانات المصطنعة وأشكالها المتعددة، إلى جانب الآليات والتقنيات المستخدمة في توليدها. ويختتم بعرض أبرز تطبيقاتها العملية في القطاعات المختلفة.

تاريخ البيانات المصطنعة

نشأ مفهوم البيانات المصطنعة في المجالات العلمية قبل أن يتبلور كمفهوم تقني مستقل يخدم الذكاء الاصطناعي وحماية الخصوصية، حيث استخدمت المحاكاة والنماذج الإحصائية لتوليد بيانات افتراضية لاختبار النماذج في مجالات مثل الفيزياء والهندسة والاقتصاد[3]، ورغم أن هذه التطبيقات لم تُشِر إلى البيانات المصطنعة بشكل مباشر، إلا أنها مهدت الطريق لفكرة إنشاء بيانات غير معتمدة مباشرة على الواقع لتنفيذ التجارب. ويعود أول استخدام منهجي للبيانات المصطنعة لأغراض حماية الخصوصية إلى عام 1993م، حين اقترح عالم الإحصاء دونالد روبن (Donald Rubin) فكرة توليد مجموعات بيانات مصطنعة بالكامل بهدف حماية سرية المعلومات الإحصائية للتعداد السكاني الأمريكي.[4]

ولكن الثورة الحقيقية جاءت في عام 2014م مع ظهور الشبكات التوليدية التنافسية (Generative Adversarial Networks – GANs)، والتي ساهمت في تحسين جودة البيانات المصطنعة، إذ يمكنها تعلم التوزيعات المعقدة للبيانات الحقيقية وخلق بيانات مصطنعة عالية الجودة، لا سيما في مجال الصور. كما أدّت المتطلبات التنظيمية المرتبطة بحماية الخصوصية، مثل اللائحة العامة لحماية البيانات (GDPR)، إلى زيادة الاهتمام بالبيانات المصطنعة كوسيلة للحفاظ على القيمة التحليلية للبيانات دون تعريض الأفراد لمخاطر الإفصاح عن بياناتهم. وفي السنوات الأخيرة، ومع انتشار الذكاء الاصطناعي التوليدي والنماذج اللغوية الكبيرة مثل جي بي تي (GPT)، ازداد دور البيانات المصطنعة أهمية كأداة أساسية لتغذية هذه النماذج خصوصاً في السياقات التي تكون فيها البيانات الحقيقية غير متاحة أو محفوفة بالمخاطر القانونية والتنظيمية.

اليوم، تُستخدم البيانات المصطنعة على نطاق واسع في مجالات عدة تشمل الذكاء الاصطناعي، وتحليل البيانات، واختبار البرمجيات، والرعاية الصحية، والتمويل، مع توقعات بأن تستمر في النمو مع ازدياد الحاجة إلى حلول تحافظ على الخصوصية وتسرّع الابتكار في مواجهة ندرة البيانات التقليدية.

مفهوم البيانات المصطنعة

تشير البيانات المصطنعة إلى البيانات المُولَّدة بهدف محاكاة الخصائص الإحصائية والأنماط والعلاقات الكامنة في البيانات الحقيقية، وتُولد هذه البيانات بعدة أساليب، بما في ذلك النماذج الإحصائية التقليدية، وتقنيات الذكاء الاصطناعي التوليدي وأساليب المحاكاة، وغيرها من تقنيات التوليد المناسبة لطبيعة البيانات وحالة الاستخدام، ويكمن الهدف الأساسي من استخدام البيانات المصطنعة في الحفاظ على القيمة التحليلية والفائدة الوظيفية لمجموعات البيانات الحقيقية، مع تعزيز حماية الخصوصية، والحد من التحيز، وتجاوز القيود المرتبطة بندرة البيانات أو صعوبة الوصول إليها.

تعتمد البيانات المصطنعة على إنشاء مجموعات بيانات جديدة كلياً، على خلاف إخفاء البيانات (Data Masking) أو إخفاء بيانات الهوية (Anonymization Data)، والذي تُستخدم فيه البيانات الحقيقية بعد حذف أو إخفاء أو تشفير القيم الحساسة كالأسماء أو المعرّفات، وهو ما يعني بقاء مخاطر إعادة التعرّف خاصة مع التطور في تقنيات تحليل البيانات والهندسة العكسية.[5] وتُعرِّف المنظمات والهيئات الدولية البيانات المصطنعة بالتركيز على خصائصها وغرضها، ومن أبرز هذه التعريفات:

- **لجنة الأمم المتحدة الاقتصادية لأوروبا (UNECE)** [6]: تُعرّف البيانات المصطنعة بأنها "بيانات تُنتَج عشوائياً (إحصائياً) بحيث تحتفظ بقيمة تحليلية عالية وفي الوقت نفسه تحقق مستويات مرتفعة من حماية السرية". وهذا يسلط الضوء على دقتها الإحصائية وقدراتها على الحفاظ على الخصوصية.

- **التحالف العالمي للجينوم والصحة (GA4GH)** [7]: يصفها بأنها "بيانات مصطنعة تحاكي عن كثب خصائص وعلاقات البيانات الحقيقية". وهذا يؤكد على محاكاة خصائص البيانات الحقيقية.

- **منظمة التعاون الاقتصادي والتنمية (OECD)** [8]: تشير إلى أن البيانات المصطنعة "هي فئة من البيانات المولَّدة اصطناعياً، على عكس البيانات التي يتم جمعها من خلال الملاحظة الواقعية". وهذا يميزها عن البيانات الحقيقية بطريقة توليدها.

- **المشرف الأوروبي لحماية البيانات (EDPS)** [9]: يصف البيانات المصطنعة بأنها "يجب أن تكون بديلاً مقبولاً إحصائياً عن البيانات الحقيقية، بحيث يعطي تحليلها نتائج متشابهة جداً لتحليل البيانات الحقيقية". بالاعتماد على التعاريف الدولية أعلاه، يمكن اعتماد التعريف التالي: "البيانات المصطنعة هي بيانات مولّدة بشكل مصطنع لمحاكاة بيانات العالم الحقيقي، تُنشأ باستخدام أساليب إحصائية أو تقنيات الذكاء الاصطناعي بحيث تمتلك نفس الخصائص الإحصائية والهيكلية للبيانات الحقيقية، ولكنها لا تحتوي على معلومات تعريفية عن أي شخص أو كيان محدد". هذا التعريف يضمن تغطية الجوانب الرئيسية، وهي: الاصطناع (عدم اشتقاقها مباشرة من أحداث حقيقية)، والمحاكاة الإحصائية الدقيقة، وانتفاء معلومات الهوية الشخصية.

الفرق بين البيانات الحقيقية والمصطنعة

على الرغم من التشابه الظاهري بين البيانات الحقيقية والبيانات المصطنعة، فإن بينهما فروقاً جوهرية، يلخص **الجدول (1)** أهم هذه الفروق.

الجدول 1: الفرق بين البيانات الحقيقية والبيانات المصطنعة.

المعيار	البيانات الحقيقية	البيانات المصطنعة
المصدر	تُجمع مباشرة من مصادر واقعية مثل التفاعلات البشرية، أو أجهزة الاستشعار، أو سجلات المعاملات.	تولَّد بواسطة خوارزميات حاسوبية، أو نماذج إحصائية، أو تقنيات الذكاء الاصطناعي التوليدي، أو أساليب المحاكاة.
الخصوصية	قد تحتوي على معلومات حساسة أو شخصية.	لا تحتوي على معلومات حساسة أو شخصية.
التكلفة والتوافر	جمعها مُكلف ومقيّد بترخيص.	قد يكون جمعها أقل كلفة وتنشأ عند الطلب وبكميات غير محدودة.
دقة التمثيل	تعكس الواقع بجميع تعقيداته، بما في ذلك التشويش والتحيّز.	قد تفشل في فهم وتمثيل الأنماط والعلاقات النادرة أو التحيّزات الكامنة في البيانات الحقيقية إذا لم تصمم وتقيم عملية التوليد بعناية.
التوازن	قد تكون غير متوازنة وغير مُمَثِّلة لبعض الفئات.	يمكن ضبطها لتغطية حالات نادرة أو فئات مُهمَّشة.

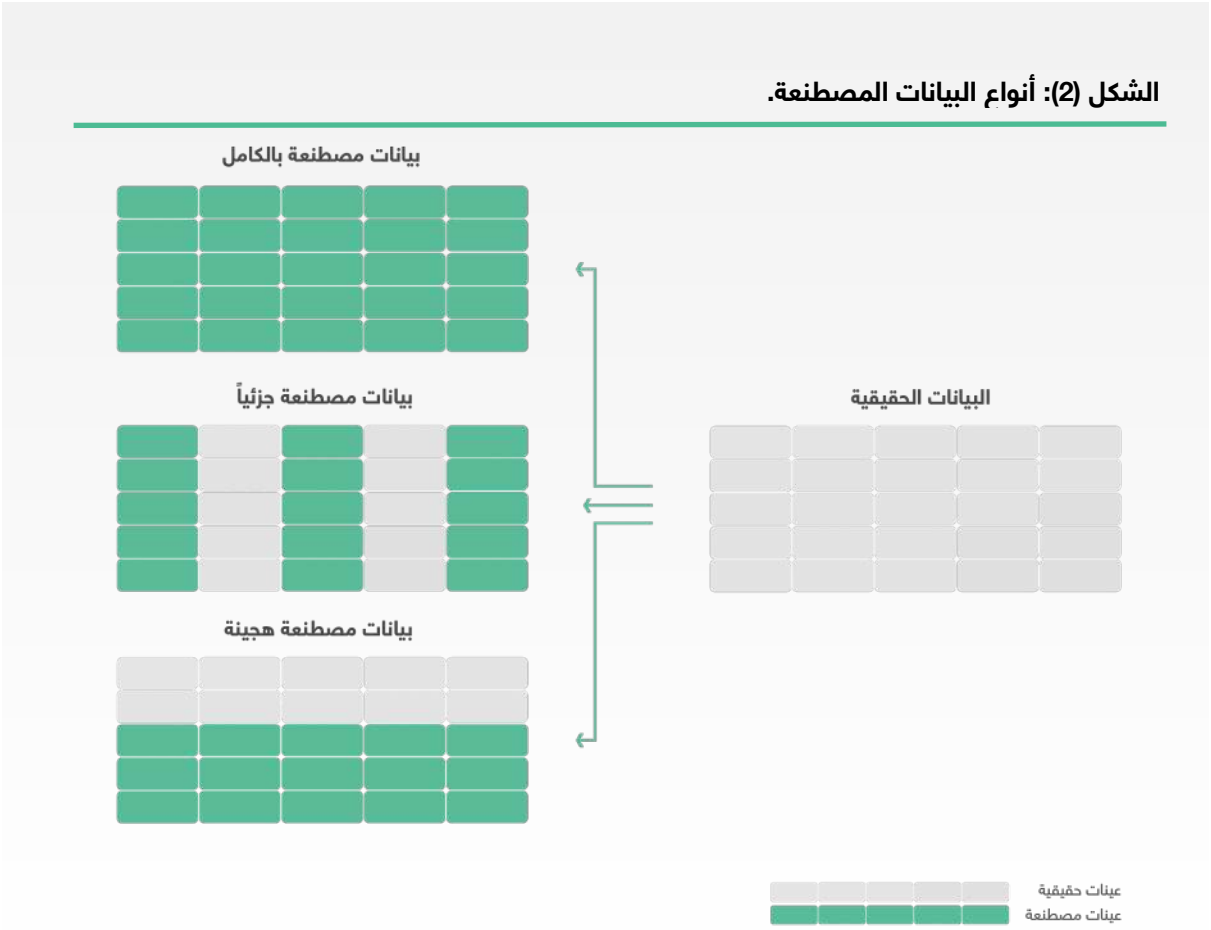
أنواع البيانات المصطنعة

يمكن تصنيف البيانات المصطنعة إلى ثلاثة أنواع وفقاً لمستوى الاصطناع في البيانات، وهي: البيانات المصطنعة بالكامل، والبيانات المصطنعة جزئياً، والبيانات المصطنعة الهجينة[10]، ويوضح **الجدول (2) والشكل (2)** هذه الأنواع بالتفصيل.

الجدول 2: أنواع البيانات المصطنعة.

	بيانات مصطنعة بالكامل	بيانات مصطنعة جزئياً	بيانات مصطنعة هجينة
التعريف	بيانات تُنشأ كلياً باستخدام الذكاء الاصطناعي أو النماذج الإحصائية أو المحاكاة، بحيث تكون جميع السجلات مصطنعة تماماً، دون أي سجلات حقيقية.	بيانات تستخدم مجموعة بيانات حقيقية كأساس، ثم تُستبدل أجزاء أو خصائص محددة منها بقيم مولدة اصطناعياً، وغالباً ما تكون خصائص حساسة وعالية الخصوصية.	بيانات تحتوي على مزيج من السجلات المصطنعة بالكامل والسجلات الحقيقية.
الهدف	إنتاج بيانات جديدة تحاكي البيانات الحقيقية بناء على تعلم الخصائص الإحصائية والارتباطات في البيانات الحقيقية.	حماية البيانات والمحافظة على الخصوصية مع الحفاظ على خصائص البيانات الحقيقية، مما يجعلها أكثر واقعية، ولكنها قد تكون أقل أماناً من البيانات المصطنعة بالكامل.	تحقيق التوازن بين الحفاظ على واقعية البيانات، وتنوعها.
الاستخدام	التطبيقات التي تتطلب درجة عالية من الخصوصية؛ لأنها تقلل من احتمالية ربط البيانات بأشخاص حقيقيين، أو لتمثيل حالات نادرة أو سيناريوهات غير موجودة.	التطبيقات التي تكون فيها البيانات الحقيقية ضرورية، مع أهمية حماية المعلومات الشخصية، أو الخصائص الحساسة مثل الاسم، والدخل، والموقع.	التطبيقات التي تتطلب زيادة البيانات الحقيقية الصغيرة الحجم بعينات مصطنعة، أو زيادة تنوع البيانات.

الشكل (2): أنواع البيانات المصطنعة.



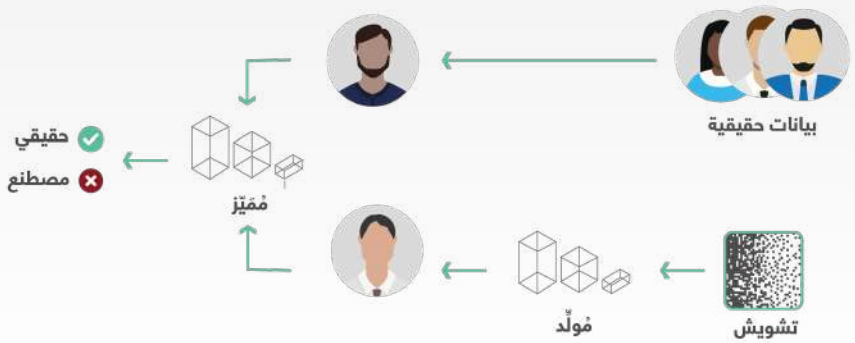
آلية إنشاء البيانات المصطنعة

يمكن توليد البيانات المصطنعة بآليات متنوعة، بدءاً من القواعد البسيطة، والطرق الإحصائية التقليدية، وحتى تقنيات الذكاء الاصطناعي التوليدي المتقدمة، ويعتمد اختيار الآلية الأنسب على طبيعة البيانات، ومتطلبات الخصوصية، ومستوى التحكم المطلوب في خصائص البيانات الناتجة، وعليه؛ يمكن تصنيف طرق توليد البيانات المصطنعة إلى ثلاثة أصناف أساسية[10]:

- **الأساليب القائمة على القواعد:** تعتمد هذه الأساليب على تحديد مجموعة من القواعد المنطقية أو الإحصائية لتتحكم في شكل البيانات المصطنعة، مثل توليد سجلات مرضى مصطنعة وفق توزيعات محددة للعمر أو الجنس أو المؤشرات الحيوية. وتُستخدم هذه الأساليب عندما تكون خصائص البيانات المراد إنتاجها واضحة وقابلة للتحكم، لكن هذه الأساليب محدودة في قدرتها على تمثيل العلاقات المعقدة بين المتغيرات.
- **النمذجة الإحصائية:** تعتمد على تحليل البيانات الحقيقية لاستخلاص التوزيعات الاحتمالية والعلاقات الارتباطية بين المتغيرات بهدف توليد بيانات جديدة تحاكي الخصائص الإحصائية للبيانات الحقيقية دون أن تكون نسخة منها. وتشمل هذه الفئة نماذج إحصائية مثل الشبكات البايزية، وسلاسل ماركوف، ونماذج الانحدار الاحتمالية.
- **نماذج الذكاء الاصطناعي التوليدي:** تمثل أكثر الأساليب تقدماً وانتشاراً في الوقت الحالي، وفيها يمكن أن تُدرَّب نماذج ذكاء اصطناعي على بيانات حقيقية لفهم بنيتها والأنماط الكامنة فيها، ثم تُستخدم النماذج المدربة في توليد بيانات مصطنعة جديدة. تُستخدم هذه الأساليب عادةً مع البيانات المعقدة مثل الصور، والفيديو، والنصوص. وغالباً ما تستخدم نماذج التعلم العميق مثل: [10]

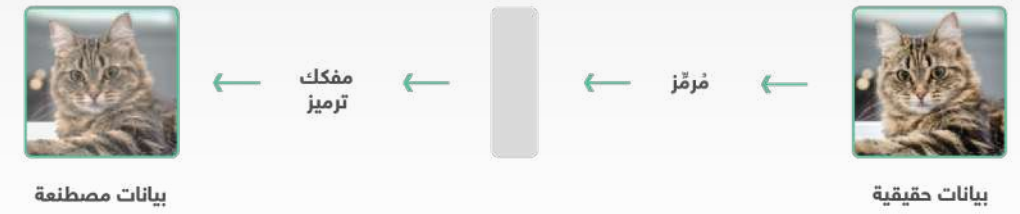
- **الشبكات التوليدية التنافسية (GANs):** تتكون هذه الشبكات من نموذجين يعملان بشكل تنافسي: مولّد يقوم بإنشاء بيانات مصطنعة، ومُميّز يحاول التفرقة بين العينات الحقيقية والمصطنعة كما يوضّح **الشكل (3)**. ومن خلال تدريب النموذجين بشكل تنافسي عبر دورات متعددة، تتحسن قدرة المولّد في إنتاج بيانات مصطنعة قد يصعب على المميز (discriminator) التفريق بينها وبين البيانات الحقيقية.

الشكل (3): آلية عمل الشبكات التوليدية التنافسية (GANs) لتوليد البيانات المصطنعة.



- **المُرمِّزات التلقائية المتغيرة (Variational Autoencoders-VAEs):** هي نماذج توليدية تقوم بضغط البيانات الحقيقية إلى تمثيل في فضاء منخفض الأبعاد، ما يساعد على استخلاص المعلومات المهمة منها، ثم تُولّد البيانات المصطنعة عبر فك ترميز هذا التمثيل. وتُستخدم هذه النماذج لتوليد بيانات متنوعة تحافظ على الخصائص المهمة للبيانات الحقيقية، كما يوضح **الشكل (4)**.

الشكل (4): آلية عمل المُرمِّزات التلقائية المتغيرة لتوليد البيانات المصطنعة.



- **النماذج الانتشارية (Diffusion Models):** تتعلم النماذج الانتشارية عبر إضافة تشويش تدريجي إلى البيانات الحقيقية، وتَعَلُّم العملية العكسية لإزالة التشويش لتوليد بيانات جديدة كما يوضح **الشكل (5)**. وقد أثبتت هذه النماذج كفاءة عالية في توليد الصور عالية الدقة.

الشكل (5): آلية عمل النماذج الانتشارية لتوليد البيانات المصطنعة.

- **النماذج اللغوية الكبيرة (Large Language Models-LLMs):** تُستخدم هذه النماذج والمبنية على المحولات (Transformers) في توليد البيانات المصطنعة النصية، وتتفوق هذه النماذج في فهم بنية اللغة وأنماطها المعقدة وتعتمد في ذلك على آلية الانتباه الذاتي لفهم العلاقات السياقية بين الوحدات اللغوية، وتوليد نصوص جديدة مترابطة لغوياً.

أشكال البيانات المصطنعة

- **النصوص المصطنعة:** تشمل النصوص التي تولّد باستخدام النماذج اللغوية الكبيرة (LLMs) مثل نماذج جي بي تي (GPT) وجروك (Grok)، ويمكن أن تأخذ شكل مقالات، أو محادثات تحاكي اللغات البشرية الطبيعية. تُستخدم هذه البيانات في تطبيقات مثل تدريب بوت المحادثة أو تحليل النصوص، وتكمن أهميتها في إمكانية استخدامها في تدريب النماذج اللغوية دون الاعتماد على بيانات حقيقية، بما يقلل مخاطر انتهاك الخصوصية، وحقوق النشر.
- **الصوتيات المصطنعة:** تتضمن الأصوات المولدة بتقنيات مثل تحويل النص إلى كلام، ويمكن أن تمثل أصواتاً بشرية متعددة اللغات واللهجات، أو أن تُستخدم في تدريب أنظمة التعرف على الكلام، وتطبيقات مساعدة المكفوفين، وتُعد هذه البيانات ذات قيمة خاصة عندما يكون جمع الأصوات الحقيقية مكلفاً أو عندما تتطلب محاكاة سيناريوهات نادرة يصعب تسجيلها واقعياً.
- **البيانات المرئية المصطنعة:** تشمل الصور والفيديوهات المصطنعة، وتولد باستخدام تقنيات متقدمة مثل الشبكات التوليدية التنافسية (GANs) ونماذج الانتشار، إلى جانب بيئات المحاكاة ثلاثية الأبعاد، ويمكن أن تُستخدم في تدريب الروبوتات والمركبات ذاتية القيادة، كما يمكن أن تتضمن توليد الوجوه البشرية لأشخاص غير موجودين لتدريب أنظمة التعرف على الوجوه. كما تُستخدم في المجال الطبي في توليد صور أشعة مصطنعة لتدريب نماذج الكشف عن الأمراض.
- **البيانات المجدولة:** هي بيانات مصطنعة منظمة في جداول، لتحاكي السجلات المالية، أو البيانات الديموغرافية. وتولد هذه البيانات مع الحفاظ على التوزيعات الإحصائية للبيانات الحقيقية وعلى العلاقات بين الخصائص المختلفة، مثل العلاقة بين العمر والدخل.

بيانات السلاسل الزمنية: تُعد بيانات السلاسل الزمنية شكلاً من البيانات يرتبط بعامل الزمن وتسلسل القيم. ويسمح هذا النوع من البيانات بتعليم الخوارزميات الأنماط الزمنية، والتنبؤ بالقيم المستقبلية، واكتشاف الشذوذ والانحرافات، ويُستخدم في تطبيقات مثل التنبؤات المالية، ومراقبة المعدات والأنظمة الصناعية، والروبوتات وإنترنت الأشياء.

- **الوسائط المتعددة:** تجمع أكثر من شكل من أشكال البيانات، مثل النصوص مع الصور أو الصوت مع الفيديو، لتوفر بيانات يمكن أن تُحاكي بيئات حقيقية متعددة الحواس، وتستخدم لتدريب النماذج متعددة الصيغ (Multimodal Models) في تطبيقات معقدة مثل السيارات الذاتية القيادة.

تطبيقات البيانات المصطنعة

- أسهمت البيانات المصطنعة بفضل قدرتها على تجاوز القيود القانونية والتقنية المرتبطة بالبيانات الحقيقية، في تبنيها في مجموعة واسعة من القطاعات والتطبيقات العملية. ومن أبرز المجالات التي يمكن أن تكون فيها البيانات المصطنعة ذات فائدة:

- **القطاع المالي والمصرفي:** يُعد القطاع المالي من أكثر القطاعات حساسية عند استخدام البيانات وأكثرها صرامة من حيث الخصوصية والامتثال التنظيمي، نظراً إلى تعامله مع معلومات شخصية ومعاملات عالية المخاطر. وقد برزت البيانات المصطنعة في دعم تطوير نماذج الذكاء الاصطناعي في مجالات إدارة المخاطر، واختبار خوارزميات التداول، والكشف عن الاحتيال وغسل الأموال، وذلك عبر توليد معاملات نادرة يصعب تمثيلها أو استخدام البيانات الحقيقية فيها. كما تُمكّن هذه البيانات المؤسسات المالية من اختبار أنظمة الامتثال ومشاركة البيانات مع شركات التقنية المالية (FinTech) دون الكشف عن بيانات العملاء الحقيقية. وفي هذا السياق، تشير استطلاعات حديثة إلى أن نحو ثلث العاملين في القطاع المالي يرون ضرورة إصدار إرشادات ومعايير تنظيمية واضحة لدعم اعتماد البيانات المصطنعة بشكل آمن ومنهجي.[11]
- **الرعاية الصحية والطب:** يُعد القطاع الصحي من أكثر المجالات استفادة من البيانات المصطنعة، نظراً إلى الحاجة إلى بيانات عالية الجودة لدعم البحث والابتكار، مقابل القيود الصارمة المرتبطة بسرية المعلومات الصحية وخصوصية المرضى. وتُستخدم البيانات المصطنعة في معالجة مشكلة ندرة البيانات الحقيقية للأمراض النادرة[12]، وتدريب نماذج التشخيص والتنبؤ، وتسهيل مشاركة البيانات بين المؤسسات البحثية دون الكشف عن معلومات شخصية. وفي المقابل، تدرك الجهات التنظيمية أن هذه الفوائد لا تخلو من التحديات، لا سيما مخاطر إعادة التعرف في الحالات النادرة أو مجموعات البيانات الصغيرة، مما يستدعي تطوير أطر تحقق صارمة توازن بين دقة البيانات وواقعيتها من جهة، وحماية خصوصية المرضى من جهة أخرى.
- **النقل والمركبات ذاتية القيادة:** تُستخدم البيانات المصطنعة على نطاق واسع في قطاع النقل، خاصة في تطوير المركبات ذاتية القيادة وأنظمة المرور الذكية. فمن خلال المحاكاة، يمكن توليد بيانات تحاكي تدفق حركة المرور، والقيادة في الظروف الجوية المختلفة، وسلوك المشاة والمركبات، كما تسمح بتدريب الخوارزميات على سيناريوهات خطرة، مثل الحوادث المفاجئة أو عبور المشاة غير المتوقع، دون الحاجة إلى اختبارات ميدانية مكلفة أو خطرة.
- **التعليم والتدريب:** تُستخدم البيانات المصطنعة لإنشاء بيئات محاكاة واقعية في مجالات التدريب عالية المخاطر أو التكاليف، مثل الطيران والجراحة والهندسة. وتتيح هذه البيئات تدريب الأفراد على سيناريوهات معقدة دون تعريضهم لمخاطر حقيقية. كما تُسهم البيانات المصطنعة في تطوير أنظمة تعليمية ذكية قادرة على التكيف مع مستوى المتعلمين، من خلال محاكاة أنماط تعلم متنوعة.
- **الإعلام والفن والترفيه:** تؤدي البيانات المصطنعة دوراً مهماً في صناعة الأفلام وألعاب الفيديو، من خلال توليد بيئات افتراضية وشخصيات ومؤثرات بصرية متقدمة، كما تفتح آفاقاً جديدة للفنانين والمبدعين، حيث تُستخدم خوارزميات الذكاء الاصطناعي لتوليد أعمال فنية، ومقطوعات موسيقية، ونصوص إبداعية اصطناعية، بما يعيد تعريف العلاقة بين التقنية والإنتاج الفني والثقافي.

ثانياً: الفرص والمخاطر

- مع تسارع الابتكارات في الذكاء الاصطناعي، برزت البيانات المصطنعة كأداة استراتيجية لتمكين الابتكار، ما يفتح آفاقاً وفرصاً واسعة، كتسهيل مشاركة البيانات بين المؤسسات دون الإخلال بالخصوصية، وتحسين تمثيل الفئات غير المُمَثلة، وتوليد بيانات لسيناريوهات لا يمكن جمعها في العالم الحقيقي، لكن هذا التوسّع يرافقه مجموعة من المخاطر والتحديات. يستعرض هذا الفصل جوانب هذه التقنية بصورة متوازنة، متناولاً الفرص الكبيرة التي توفرها من تعزيز الخصوصية وإتاحة التعاون، وتمكين نماذج تعلم الآلة من التدريب في بيئات آمنة، إلى جانب المخاطر المرتبطة بجودة البيانات، والتحيز، والأمن والحوكمة.

الفرص والإمكانات

- توفر البيانات المصطنعة مجموعة واسعة من الفرص والإمكانات التي شأنها تسريع وتيرة الابتكار في مجال الذكاء الاصطناعي مع تعزيز الامتثال التنظيمي والممارسات الأخلاقية، ومن أبرز هذه الفرص:[9] [13]

- **حماية الخصوصية:** تشبه البيانات المصطنعة البيانات الحقيقية من حيث الخصائص الإحصائية والهيكلية، دون أن تحتوي على معلومات شخصية أو سجلات يمكن ربطها بأفراد أو كيانات، وهي بذلك تعمل كآلية متقدمة لإخفاء الهوية، ما يجعلها إحدى تقنيات تعزيز الخصوصية (Privacy-Enhancing Technologies)، إذ تحد من احتمالات تسرب البيانات والوصول غير المصرح به إلى المعلومات الحساسة.
- **الامتثال التنظيمي:** تعد البيانات المصطنعة أداة مهمة للامتثال للوائح حماية البيانات، مثل اللائحة العامة لحماية البيانات (GDPR) في الاتحاد الأوروبي، ولوائح القطاعات الخاضعة لتنظيمات صارمة، مثل الرعاية الصحية، والتمويل، والطاقة، والدفاع، والتي تفرض قيوداً شديدة على استخدام البيانات ومشاركتها. حيث تمكّن البيانات المصطنعة هذه الجهات من إجراء التحليلات والأبحاث، أو مشاركة البيانات مع أطراف أخرى، دون خرق المتطلبات القانونية أو الأخلاقية، كما تساعد على تجنب إشكالات الملكية الفكرية وحقوق النشر المرتبطة باستخدام بيانات حقيقية.
- **سد فجوات البيانات ومعالجة ندرتها:** تعاني عدد من تطبيقات الذكاء الاصطناعي من نقص البيانات وعدم كفايتها لحالات الاستخدام المطلوبة، وتبرز هذه المشكلة بوضوح في السيناريوهات النادرة أو الحساسة، مثل الأمراض النادرة، أو أنماط الاحتيال غير الشائعة، أو الحوادث الخطرة في أنظمة القيادة الذاتية، كما تسمح بمحاكاة أحداث يصعب أو يستحيل جمع بيانات حقيقية عنها، مثل السيناريوهات العسكرية عالية الخطورة، أو السيناريوهات المستقبلية[13]. كما تتيح البيانات المصطنعة توليد كميات كبيرة وغير محدودة تقريباً من البيانات الموسومة عند الطلب، مما يسهم في معالجة ندرة البيانات، ويحسن أداء نماذج تعلم الآلة.
- **تحسين جودة البيانات وتقليل التحيز:** غالباً ما تكون البيانات الحقيقية فوضوية وغير مكتملة أو غير متوازنة من حيث تمثيل الفئات المختلفة، كما أنها قد تعكس تحيزات اجتماعية. يمكن أن تستخدم البيانات المصطنعة لمعالجة هذه الإشكاليات عبر موازنة الفئات غير المُمَثلة بشكل كافٍ، وتوليد عينات إضافية للفئات النادرة، وزيادة مجموعات البيانات الصغيرة، مما يساعد على تقليل التحيز والتمييز في الخوارزميات، ويُسهم في بناء نماذج ذكاء اصطناعي أكثر عدالة، ويقلل أثر التحيزات الكامنة في البيانات الحقيقية، ويحد من مخاطر التمييز الخوارزمي.
- **مشاركة البيانات وتمكين التعاون:** تدعم البيانات المصطنعة مشاركة البيانات بين الإدارات داخل المؤسسة الواحدة، ومع الأطراف الخارجية سواء كانت بين شركات تقنية، أو باحثين، أو مزودي خدمات، كما قد تسمح بتجاوز القيود القانونية المرتبطة بنقل البيانات عبر الحدود دون الكشف عن بيانات حساسة أو الإخلال بقوانين الخصوصية، مما يسهّل التعاون، والتطوير المشترك، واختبار النماذج بشكل جماعي.
- **تسريع الابتكار وخفض التكاليف:** تُعد عملية جمع البيانات الحقيقية وتسميتها مكلفة وتستغرق وقتاً طويلاً، وتوفر البيانات المصطنعة بديلاً لهذه العملية قد يسمح بتوليد بيانات عالية الجودة بسرعة وبتكلفة أقل. وهذا بدوره يسرع من دورة تطوير

نماذج الذكاء الاصطناعي، ويدعم بناء النماذج الأولية والتجريب السريع. علاوة على ذلك، فإنه يضفي طابعاً ديمقراطياً على الوصول إلى البيانات، مما يمكّن المؤسسات الصغيرة والباحثين الذين قد لا يملكون الموارد اللازمة للوصول إلى البيانات.

المخاطر والتحديات

على الرغم من الفرص والإمكانات الكبيرة التي توفرها البيانات المصطنعة، فإن اعتمادها دون ضوابط واضحة وأطر حوكمة قد تترتب عليه مجموعة من المخاطر التقنية والقانونية والأخلاقية، ومن أبرزها:

- **مخاطر الخصوصية وإعادة التعرف:** على الرغم من أن البيانات المصطنعة تستخدم لتقليل مخاطر انتهاك الخصوصية، فإن هذا الخطر لا يختفي بالكامل، إذ إن البيانات المصطنعة قد تبقى عرضة لإعادة التعرف[5]، لا سيما إذا كانت نماذج التوليد قد حفظت أنماطاً دقيقة أو نادرة من البيانات الحقيقية. لذا تؤكد الدراسات على ضرورة إجراء تقييمات كمية لمخاطر إعادة التعرف لقياس مقدار المعلومات التي يمكن استنتاجها من البيانات المصطنعة. وفي السياقات عالية الحساسية، مثل الدفاع أو الأمن القومي، يحذّر تقرير للأمم المتحدة من مخاطر إضافية، تشمل احتمال كشف معلومات أمنية حساسة.[13]
- **ضعف الواقعية ومحدودية التعميم:** من أبرز التحديات التي تواجه البيانات المصطنعة احتمال عدم تمثيلها الدقيق لتعقيدات الواقع. فقد تفشل نماذج التوليد في التقاط الحالات الشاذة، أو التفاعلات النادرة، أو التفاصيل السياقية الدقيقة الموجودة في البيانات الحقيقية، مما يؤدي إلى تدريب نماذج ذكاء اصطناعي تؤدي أداءً ضعيفاً عند التعامل مع العالم الحقيقي. وقد يؤدي هذا القصور أحياناً إلى إدخال تحيزات جديدة غير مقصودة في بيانات التدريب، أو إلى تجاهل أنماط حرجة، الأمر الذي يستدعي إضافة مراحل تحقق واختبار إضافية، تزيد من تعقيد مرحلة معالجة البيانات وتطيل دورة تطوير النماذج.
- **تدهور النماذج وانهيارها بمرور الوقت:** تشير عدد من الدراسات إلى خطر ما يُعرف بانهيار النماذج (Model Collapse)[14] ، والذي يحدث عند تدريب النماذج بشكل متكرر على بيانات مصطنعة مولَّدة بواسطة نماذج ذكاء اصطناعي أخرى. وفي هذه الحالة، قد تفقد النماذج تدريجياً تنوع البيانات وثراء التوزيعات الأصلية، ما يؤدي إلى تضخيم العيوب، وانخفاض جودة التعميم، بل وابتعاد النماذج عن أنماط السلوك البشري أو الواقعي.
- **إعادة إنتاج التحيز:** على الرغم من أن البيانات المصطنعة تُستخدم أحياناً لتقليل التحيز، إلا أنها قد تعيد إنتاج التحيزات الموجودة في البيانات الحقيقية إذا لم تُصمَّم بعناية، فإذا كانت البيانات الحقيقية غير متوازنة أو منحازة ديموغرافياً، فإن النماذج التوليدية قد تعكس هذه الاختلالات أو حتى تضخمها عند توليد البيانات المصطنعة، ويستلزم التخفيف من هذا الخطر استخدام مصادر بيانات متنوعة، وإدخال ضوابط صريحة لإعادة التوازن، والتحقق من تمثيل الفئات المختلفة، إلى جانب اختبار النماذج الناتجة على بيانات حقيقية مستقلة.

ومن المهم التمييز بين البيانات المصطنعة المستخدمة لأغراض التدريب والاختبار والتحليل، وبين المحتوى المولد بالذكاء الاصطناعي؛ إذ تختلف المخاطر المرتبطة بكل منهما. فالتزييف العميق وانتحال الهوية والتضليل ترتبط بصورة أساسية بالمحتوى المولد، بينما تتركز مخاطر البيانات المصطنعة في جوانب مثل جودة البيانات والتحيز والخصوصية وإمكانية إعادة التعرف.

ثالثاً: المشهد العالمي

تشهد تقنيات البيانات المصطنعة تطوراً متسارعاً، يقابله اهتمام على المستوى العالمي من قبل الحكومات والمنظمات الدولية، نظراً إلى دورها المتنامي في دعم الابتكار، وتمكين الاستخدام الآمن للبيانات. يقدم هذا الفصل مدخلاً عاماً لحوكمة البيانات المصطنعة، مع التركيز على أبعادها التنظيمية والتشريعية الرئيسية. كما يستعرض أبرز المبادرات الدولية والسياسات الوطنية ذوات الصلة، بهدف استخلاص الدروس المستفادة وتحديد أفضل الممارسات العالمية التي أثبتت فعاليتها في سياقاتها المحلية. ويسلط الفصل الضوء كذلك على الاتجاهات التنظيمية الناشئة والفجوات القائمة، وذلك من خلال مراجعة منهجية للسياسات والمعايير المعتمدة على الصعيد الدولي.

حوكمة البيانات المصطنعة

تتضمن الحوكمة في سياق البيانات المصطنعة مجموعة الأطر والقواعد والعمليات التشغيلية والمسؤوليات المحدَّدة التي توجه إنشاء هذه البيانات واستخدامها وإدارتها والإشراف عليها[15]. وتنشأ الحاجة إلى مثل هذه الحوكمة لضمان الامتثال للمعايير الأخلاقية والقانونية، وتعزيز الشفافية والمسؤولية في استخدام البيانات المصطنعة، إضافة إلى تقليل المخاطر المرتبطة بانتهاكات الخصوصية والتحيزات المحتملة في النماذج المستندة إلى هذه البيانات.

إلا أن الحوكمة تواجه تحدياً إضافياً يتمثل في الدور المزدوج الذي تؤديه البيانات المصطنعة. فبينما تتطلب هذه البيانات هياكل حوكمة قوية لإدارة إنشائها واستخدامها، فإنها تُستخدم في الوقت نفسه كأداة للحوكمة والتنظيم في مجالات أخرى من الذكاء الاصطناعي. على سبيل المثال، يجري التحقيق في البيانات المصطنعة من حيث فائدتها في تدقيق أنظمة الذكاء الاصطناعي من حيث الإنصاف والتحيز، واستخدامها في التدريب العدائي لتعزيز متانة وسلامة نماذج تعلم الآلة. وتعني هذه الازدواجية أن البيانات المصطنعة لديها القدرة على "إضعاف نماذج الحوكمة الحالية" إذا كانت حوكمتها الخاصة غير كافية، لكنها في المقابل تتيح فرصاً مهمة للرقابة والتنظيم، حيث يمكن توظيفها لتحسين أمان أنظمة الذكاء الاصطناعي وعدالتها وكفاءتها، مما قد يحسن حوكمة أنظمة الذكاء الاصطناعي الأخرى. تمثل هذه الخاصية المزدوجة تحدياً فريداً للحوكمة: كيف يمكن حوكمة تقنية هي في حد ذاتها أداة محتملة للحوكمة؟ وهذا يتطلب توخي الحذر والدراسة المتأنية لتجنب المخاطر المحتملة. [16]

الأبعاد التنظيمية

تُعَدّ البيانات المصطنعة أداة مهمة لتجاوز الإشكالات التقنية والقانونية التي تفرضها البيانات الحقيقية، إلا أن توليد هذه البيانات واستخدامها يثير مجموعة من القضايا التنظيمية التي تتطلب حوكمة متوازنة، تجمع بين حماية الخصوصية، وضمان جودة البيانات، وتعزيز الشفافية، وتحقيق العدالة.

حماية الخصوصية

تعد حماية هوية الأفراد أحد الدوافع الأساسية لتبني البيانات المصطنعة، إلا أن هذه البيانات لا تضمن الخصوصية بشكل كامل، ما يخلق خطراً يُعرف بإعادة التعرف، ويوصف بأنه إمكانية ربط بيانات مصطنعة بأفراد محددين، ويعتبر هذا الخطر أحد أبرز التحديات القانونية المرتبطة بالبيانات المصطنعة، وتستلزم مواجهة هذا الخطر:

- **اختبارات إعادة التعرف:** تساعد هذه الاختبارات على تقييم خطر إعادة التعرف قبل نشر البيانات، وتشمل اختبارات مثل اختبار الفصل (Singling out) أي القدرة على تمييز سجل يخصّ فرداً محدداً، والربط (Linkability) أي إمكانية ربط سجلين أو أكثر للشخص نفسه في مجموعات بيانات مختلفة، والاستنتاج (Inference) أي استنتاج معلومات جديدة عن فرد من بيانات تبدو مجهولة، كما تستخدم طرق أخرى مثل تحليل التشابه الإحصائي بين البيانات الحقيقية والمصطنعة.

- **تعزيز خصوصية البيانات:** وذلك باستخدام تقنيات مثل الخصوصية التباينية (Differential Privacy) وهي إطار رياضي يضيف تشويشاً منظماً إلى البيانات مع الحفاظ على البنية والخصائص الإحصائية العامة للبيانات الحقيقية، ويحدد مستوى الحماية المطلوبة بمعاملين هما إبسيلون (ε) ودلتا (δ).
- **منع أو تجريم محاولات إعادة التعرف:** إما بوضع عقوبات أو غرامات، أو بفرض تعهدات والتزامات على مستخدمي البيانات بعدم محاولة تحديد هوية شخص من البيانات المصطنعة.

جودة البيانات

تعد جودة البيانات المصطنعة خاصية مهمة لضمان سلامة القرارات المستندة إلى هذه البيانات. فضعف الجودة قد يؤدي إلى نتائج مضلِّلة، لذا ينبغي أن تضمن سياسات الحوكمة جودة البيانات، والتي تتضمن مفهومين أساسيين:

- **المصداقية (Fidelity):** وتمثل مدى احتفاظ البيانات المصطنعة بالتوزيعات والعلاقات الموجودة في البيانات الحقيقية، وتستخدم في ذلك مقاييس التشابه الإحصائي مثل: تباعد كولباك–ليبْلِر (KL Divergence).
- **المنفعة (Utility):** وتمثل القيمة التحليلية للبيانات المصطنعة، وتستخدم لقياسها مقاييس مثل الدقة.

وتبرز معضلة أمام الحوكمة عند محاولة الموازنة بين جودة البيانات وحماية الخصوصية، فزيادة تدابير الخصوصية مثل إضافة التشويش، يؤدي إلى انخفاض مصداقية البيانات، وبالمثل، كلما كانت قدرة البيانات المصطنعة على محاكاة الخصائص الإحصائية للبيانات الحقيقية أكبر، ارتفعت معها احتمالية المخاطر المتعلقة بالخصوصية، هذه المعضلة تستدعي اعتماد مقاييس مزدوجة تجمع ما بين الجودة والخصوصية، والموازنة بين هذه المقاييس. ويُعد تحقيق التوازن المناسب بين هذين البعدين أمراً سياقياً يعتمد على طبيعة الاستخدام، ويتطلب قرارات حوكمة واضحة وتقييمات مخاطر مستمرة ولا ينبغي تقييم جودة البيانات المصطنعة اعتماداً على المقاييس الفنية وحدها، بل ينبغي ربط التقييم بطبيعة البيانات وحالة الاستخدام والهدف من التوليد، مع إمكانية الاستعانة بالمراجعة البشرية عند الحاجة، وتحديد معايير التقييم مسبقاً.

العدالة وتخفيف التحيّز

يُعدّ التحيّز قضية مهمة بالنسبة للبيانات المصطنعة، فمن جهة يمكن أن تُستخدم هذه البيانات كأداة لتصحيح التحيزات الموجودة في البيانات الحقيقية كتلك المتعلقة بالجنس أو العِرق، ومن جهة أخرى، قد تعيد البيانات المصطنعة إنتاج التحيزات الموجودة في البيانات الحقيقية، خصوصاً إذا لم تُضبط آليات التوليد بعناية. [13]

ويكمن التحدي في ضبط التحيز في البيانات المصطنعة، في أن مسؤولية الحوكمة تتحول من حوكمة البيانات إلى حوكمة النماذج التوليدية نفسها، باعتبارها مصدراً محتملاً لهذه التحيزات، وتتطلب الحوكمة السليمة رصد التحيز من خلال مقارنة الخصائص التوزيعية للبيانات المصطنعة بالبيانات الحقيقية، كما ينبغي أن تلتزم المؤسسات بإجراء اختبارات كشف التحيز (Bias Detection)، وينبغي تقييم عدالة البيانات المصطنعة في ضوء حالة الاستخدام والفئات المتأثرة والهدف من استخدامها، ولا يكفي تطابق التوزيعات الإحصائية مع البيانات الأصلية للحكم على عدالتها، إذ قد تؤدي محاكاة البيانات الأصلية إلى إعادة إنتاج التحيزات الكامنة فيها. ويشمل ذلك التدقيق في قدرة النماذج التوليدية على التعامل مع القيم الشاذة (Outliers)، ويرتبط تحليل القيم الشاذة (Outliers) بصورة أساسية بتقييم جودة البيانات وقدرتها على تمثيل الحالات النادرة والمتطرفة، ويمكن أن يسهم بصورة غير مباشرة في تحليل التحيز عندما تكون هذه القيم مرتبطة بفئات أو حالات محددة. ومن التدابير الممكنة أيضاً تطوير أدوات متخصصة لرصد التحيّزات الخوارزمية وتخفيفها قبل اعتماد البيانات المصطنعة، وتُعد هذه الاختبارات خطوة أساسية لضمان أن تكون البيانات المصطنعة دقيقة وعادلة وقابلة للاستخدام.

الشفافية

تُعد الشفافية حجر الأساس في بناء الثقة العامة في استخدام البيانات المصطنعة، إذ تهدف إلى ضمان النزاهة وتعزيز مصداقية القرارات والأبحاث التي تعتمد عليها. ويؤكد الخبراء على أهمية تبنّي سياسات واضحة تضمن معرفة المستخدمين والجهات الرقابية بطبيعة هذه البيانات وكيفية إنتاجها واستخدامها. ويشمل ذلك تسمية البيانات المصطنعة عند نشرها والإفصاح عن أنها مولَّدة اصطناعياً، على نحو مماثل لما تفرضه بعض التشريعات على المحتوى المُولَّد بالذكاء الاصطناعي لمكافحة التزييف العميق، إلى جانب توفير وثائق تعريفية أو "بطاقات بيانات" توضّح مصدر البيانات الحقيقية، وحدود الاستخدام، والمنهجيات المتبعة في التوليد. كما تمتد الشفافية إلى الإفصاح عند استخدام البيانات المصطنعة في تدريب نماذج الذكاء الاصطناعي أو في دعم قرارات مؤثرة، فضلاً عن الاحتفاظ بسجلات تدقيق تتضمن مؤشرات الجودة والتحيز، بما يمكّن الجهات الرقابية من مراجعة عمليات التوليد وتقييمها، ويعزّز المساءلة والحوكمة.

الجدول 3: أبعاد حوكمة البيانات المصطنعة والأدوات المستخدمة في التنظيم.

المحور	الهدف	الأدوات والمقاييس
الخصوصية	تقليل مخاطر إعادة التعرف وضمان الامتثال لقوانين حماية البيانات.	• اختبارات إعادة التعرف مثل الفصل، والربط، والاستنتاج. • الخصوصية التباينية. • الحظر القانوني لمحاولات إعادة التعرف.
الجودة	ضمان مصداقية البيانات المصطنعة ومنفعتها العملية.	• مقاييس التشابه الإحصائي مثل تباعد كولباك–ليبْلِر (KL Divergence)، ومقياس جنسن–شانون (JSD). • مقاييس الأداء مثل الدقة، والاستدعاء، والإحكام.
العدالة	تقليل التحيز وضمان تمثيل عادل لجميع الفئات.	• اختبارات كشف التحيز. • مقارنة الخصائص التوزيعية مع البيانات الحقيقية.
الشفافية	تعزيز الثقة والمساءلة عبر الإفصاح والتتبع.	• بطاقة البيانات المصطنعة. • تسمية البيانات. • سجلات التدقيق.

التجارب الدولية

على الرغم من الانتشار الواسع لاستخدام البيانات المصطنعة، إلا أن حوكمة هذا النوع من البيانات لم تصل بعد إلى مستوى النضج القانوني. فيما يلي أبرز التوجهات والمبادرات التنظيمية في عدد من الدول والمنظمات، التي تسعى إلى بناء حوكمة أكثر وضوحاً لهذا النوع من البيانات.

الاتحاد الأوروبي

لا يمتلك الاتحاد الأوروبي تشريعاً مخصصاً ينظم البيانات المصطنعة، إلا أن المفهوم حاضر في عدد من التشريعات الأوروبية المعنيّة بحوكمة البيانات والذكاء الاصطناعي، ومن أهمها اللائحة العامة لحماية البيانات (GDPR)[1]، وقانون الذكاء الاصطناعي (AI Act)[17] إلى جانب قانون حوكمة البيانات (Data Governance Act)[18].

تُعنى اللائحة العامة لحماية البيانات (GDPR) بتنظيم معالجة البيانات الشخصية، وتُعرّف البيانات الشخصية في المادة 4 (1) بأنها "أي معلومات تتعلق بشخص مُعرَّف أو قابل للتعريف"[1]، وعلى الرغم من أن اللائحة تطبق على البيانات الشخصية الحقيقية لا البيانات المصطنعة، إلا أنها أشارت إلى أنها تستبعد من نطاق تطبيقها البيانات مجهولة الهوية، والتي تتعذّر عملياً إعادة تحديد الأفراد منها، أما إذا احتوت أنماطاً أو خصائص تسمح بإعادة التعرف فإنها تُعامل كبيانات شخصية تنطبق عليها أحكام اللائحة، ولكن يظل هذا التصنيف غامضاً من الناحية العملية وما إذا كان يشمل البيانات المصطنعة أم لا. في هذا السياق أشار المشرف الأوروبي على حماية البيانات (EDPS) وهو الجهة المسؤولة عن تطبيق اللائحة على البيانات المصطنعة، إلى أن مستخدمي البيانات المصطنعة يتعين عليهم إجراء تقييمات منهجية للخصوصية للتحقق من خلو البيانات المصطنعة من أنماط أو إشارات قد تسمح بإعادة التعرف على البيانات الحقيقية، وتحديد طبيعة المعلومات التي قد تُستخلص في حال وقوع ذلك.[9]

أما قانون الذكاء الاصطناعي(AI Act)[17]، الذي يُعد أول إطار تشريعي شامل للذكاء الاصطناعي، فقد تضمن في مواده إشارات مباشرة إلى البيانات المصطنعة، واعتبرها أداة للامتثال التنظيمي، خصوصاً فيما يتعلق بجودة البيانات، والشفافية، والحد من التحيّز. ومن أبرز هذه الإشارات:

- **المادة 10 (5)(أ):** التي تنص على أن الأنظمة عالية المخاطر يجب أن تُدرَّب على بيانات عالية الجودة، وتجيز استخدام البيانات المصطنعة لتصحيح التحيّز، أو عندما تكون البيانات الحقيقية غير كافية.
- **المادة 50 (2):** التي تفرض التزامات تتعلق بالشفافية على أنظمة الذكاء الاصطناعي التوليدي، بما في ذلك إلزامها بالإفصاح عن أن المحتوى الناتج (نصوص، صور، صوت، فيديو) قد تم توليده بشكل اصطناعي.
- **المادة 59 (1)(ب):** التي تشير إلى البيانات المصطنعة في سياق بيئات الاختبار التنظيمية (AI Sandboxes)، حيث تُعامل جنباً إلى جنب مع البيانات مخفية الهوية كبدائل لاستخدام البيانات الشخصية الحقيقية في التجريب والاختبار.

من ناحية أخرى فإن قانون حوكمة البيانات (Data Governance Act)[18] الذي يهدف إلى تنظيم مشاركة البيانات العامة الخاضعة لحماية خاصة مثل الأسرار التجارية، والملكية الفكرية، والبيانات الشخصية، قد أشار في الديباجة 7 (7 Recital) إلى البيانات المصطنعة كتقنية لحماية الخصوصية تسمح بمشاركة البيانات بشكل آمن.

وبناءً على ما سبق، يمكن القول إن الإطار التشريعي الأوروبي يعترف بالبيانات المصطنعة ولكن ليس بوصفها فئة بيانات تخضع لاعتبارات قانونية مستقلة، بل باعتبارها أداة لدعم الامتثال لقوانين حماية البيانات والذكاء الاصطناعي وتعزيز حماية الخصوصية.

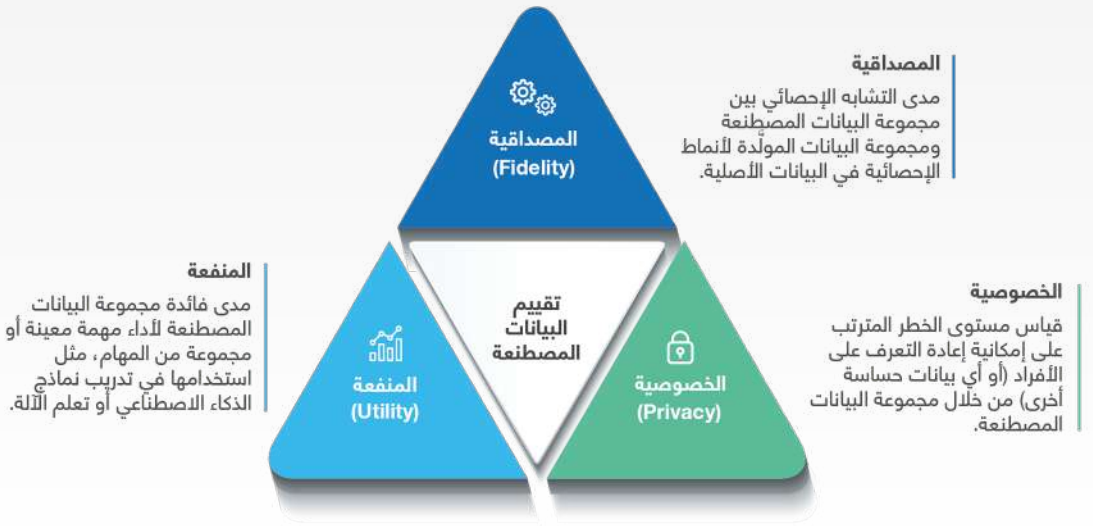
المملكة المتحدة

ترتكز المملكة المتحدة في تنظيم البيانات الحقيقية الشخصية على قانون حماية البيانات (Data Protection Act) [19] واللائحة العامة البريطانية لحماية البيانات (UK GDPR) [20]، التي تعرف البيانات الشخصية بأنها "أي معلومات تتعلق بشخص طبيعي مُعرَّف أو قابل للتعرّف". ورغم أن التشريعات لا تذكر البيانات المصطنعة، إلا أن تصنيفها القانوني يعتمد على قابلية التعرف؛ فإذا أمكن ربطها بهوية شخص حقيقي عُدَّت بيانات شخصية، أما إذا كان خطر إعادة التعرف ضئيلاً فتُعامل كبيانات مجهولة. كما يحتوي قانون حماية البيانات على المادتين (170–171) اللتين تجرّمان أي محاولة متعمّدة لإعادة التعرف على بيانات أُزيلت هويتها، وهو ما قد ينطبق على البيانات المصطنعة إذا استُخدمت كآلية لإخفاء الهوية. ويبرز في المملكة المتحدة دور مكتب مفوض المعلومات الذي يضع الإرشادات العملية التي تفسر الأطر القانونية السابقة، والتي أدرجت البيانات المصطنعة ضمن تقنيات تعزيز الخصوصية في تحديث عام 2022م[20]، كما قدمت إطاراً منهجياً لتقييم قابلية التعرف، يستند إلى اختبارات عملية مثل إمكانية الفصل، والربط، والاستدلال، واختبار المتسلل المحفَّز (Motivated Intruder Test) الذي يفترض وجود خصم يمتلك جميع الوسائل لمحاولة إعادة التعرف على الأفراد من خلال البيانات.

على مستوى التطبيق العملي، تستخدم الحكومة البريطانية القطاع العام كمختبر حي لاختبار البيانات المصطنعة، على سبيل المثال يوفر مكتب الإحصاءات الوطنية بيانات مصطنعة تجريبية لتمكين الباحثين من تطوير البرمجيات وتحليل البيانات قبل الحصول على الموافقات اللازمة للوصول إلى البيانات الحقيقية، ما يقلّص زمن وتكلفة الوصول إلى البيانات مع الإبقاء على الخصوصية[21]. وبالمثل، أطلقت هيئة الخدمات الصحية الوطنية مشروعاً تجريبياً لتوفير 200 مجموعة بيانات مصطنعة في القطاع الصحي لحل مشكلة الوصول إلى البيانات الصحية.[22]

أما في القطاع المالي فتُعد المملكة المتحدة من الدول الرائدة التي اختبرت تطبيقات البيانات المصطنعة ضمن البيئات التنظيمية التجريبية التي تشرف عليها هيئة السلوك المالي[11]. وقد استُخدمت هذه البيانات في حالات مثل مكافحة غسل الأموال والكشف عن الاحتيال، بما يتيح اختبار النماذج دون المساس ببيانات العملاء. كما ساهمت الهيئة في تطوير إطار "مثلث الجودة" الذي يوازن بين المنفعة والمصداقية والخصوصية (Fidelity, Utility, and Privacy)[11] كما يوضح **الشكل (6)**، وتعمل الهيئة مع جهات (مثل مكتب مفوض المعلومات) على تحقيق هذا التوازن عند التعامل مع البيانات.

الشكل (6): إطار مثلث الجودة.



كما أوصت هيئة السلوك المالي في تقريرها بدمج عمليات التوليد والتحقق ضمن أطر إدارة مخاطر النماذج، بحيث تكون مسؤولية التحقق من جودة البيانات جزءاً من عملية المراجعة المؤسسية. كما أكدت على ضرورة أن تُراجع لجان المخاطر في المؤسسات المالية مصدر البيانات التدريبية لضمان المساءلة والشفافية، مع ضرورة وضع معايير موحّدة للجودة تضمن احتفاظ البيانات المصطنعة بالعلاقات الإحصائية الجوهرية، وإدراج التزامات تعاقدية واضحة عند مشاركة البيانات المصطنعة مع أطراف خارجية، لحظر أي محاولات لإعادة التعرف، وتبنّي المراقبة المستمرة عبر أنظمة تقيس المصداقية بشكل دوري، وتصدر تنبيهات عند انحراف البيانات المصطنعة عن الحدود التنظيمية المقبولة، وتحديد الملكية والمسؤولية عن جودة النماذج المولدة.

وبصورة عامة، تُدار حوكمة البيانات المصطنعة في المملكة المتحدة عبر الإرشادات العملية والبيئات التجريبية والأبحاث المشتركة، رغم غياب القوانين واللوائح التشريعية، هذا النهج العملي يعكس اعترافاً بالبيانات المصطنعة بوصفها أداة امتثال فعّالة ضمن منظومة حماية الخصوصية، ويقدّم نموذجاً مرناً يمكن أن تستفيد منه قطاعات ودول أخرى.

الولايات المتحدة الأمريكية

لا تمتلك الولايات المتحدة حتى الآن قانوناً فدرالياً شاملاً لحماية البيانات أو لتنظيم البيانات المصطنعة على وجه التحديد. فالإطار القانوني الأمريكي يقوم على مزيج من القوانين القطاعية مثل قانون قابلية النقل والمساءلة للتأمين الصحي (HIPAA)، والتشريعات على مستوى الولايات، ما يعني غياب موقف تنظيمي وطني موحّد تجاه البيانات المصطنعة. إلا أن بعض المقترحات التشريعية الفدرالية أشارت إليها في سياق تقنيات تعزيز الخصوصية. مثل مشروع "قانون تعزيز تقنيات الخصوصية الرقمية" الذي قُدّم للكونغرس في عام 2022م، إلا أن هذا المشروع لم يُقرّ ليصبح قانوناً نافذاً.[23]

على مستوى الولايات الأمريكية، برزت ولاية يوتا كأول ولاية تُدرج البيانات المصطنعة ضمن تشريعاتها من خلال تعديلات عام 2024م على قانون حماية خصوصية المستهلك (Utah Consumer Privacy Act – UCPA)[24] ، والذي عرّفها بأنها "بيانات مولَّدة بخوارزميات أو نماذج إحصائية ولا تحتوي على معلومات شخصية". ويُعامل القانون هذه البيانات على أنها بيانات مُزالة الهوية، وهو ما يعني أنها لا تخضع لعدد من الالتزامات الخاصة بالبيانات الشخصية. ويُشترط ضمن هذا الإطار أن تتخذ الجهة المالكة لهذه البيانات إجراءات لمنع ربط هذه البيانات مع أشخاص معرفين، وأن تُلزم المتلقين لها التعهد بالامتثال لذلك، كما يَعتبر القانون أي محاولة للكشف عن الهوية من هذه البيانات مخالفة تعاقدية وتنظيمية. أما في ولاية كاليفورنيا، فقد أُقر مشروع قانون "الشفافية في استخدام البيانات في أنظمة الذكاء الاصطناعي" الذي يلزم مطوري أنظمة الذكاء الاصطناعي بالإفصاح عما إذا كانوا يستخدمون بيانات مصطنعة، إلى جانب تفاصيل أخرى حول مصادر وخصائص البيانات المستخدمة[25]. وقد قُدّمت في عدة ولايات مثل نيويورك وواشنطن مشاريع لقوانين مشابهة. ورغم أنها لم تُقر بصفة نهائية بعد، إلا أنها تعكس توجهاً تشريعياً نحو إلزامية الإفصاح عن استخدام البيانات المصطنعة وضمان نزاهتها.

كما يؤدي المعهد الوطني للمعايير والتقنية (NIST) دوراً مهماً في وضع المصطلحات التقنية وصياغة الأطر الخاصة بالمحتوى المصطنع والبيانات المصطنعة، ومن أبرز إصداراته:

- **إطار عمل إدارة مخاطر الذكاء الاصطناعي (NIST AI RMF 1.0):** إطار طوعي لإدارة مخاطر الذكاء الاصطناعي، يتناول مخاطر المحتوى المصطنع، بما في ذلك البيانات المصطنعة والتزييف العميق، ويوفّر مؤشرات للمؤسسات حول كيفية تعزيز الشفافية، وتقليل التحيّز، ويُستخدم كمرجع للمؤسسات الحكومية والخاصة لتعزيز الاستخدام الآمن للذكاء الاصطناعي.[26]
- **الحد من المخاطر التي يفرضها المحتوى الاصطناعي (NIST AI 100-4):** تقرير تقني يركّز على مخاطر التزييف العميق، ويقدّم مجموعة من الأساليب لتقليل هذه المخاطر، مثل آليات التوثيق وإثبات المصدر، والعلامات المائية، والتوثيق المدمج في البيانات. يوصي التقرير أيضاً بوضع معايير موحّدة لتصنيف وتسمية المحتوى المصطنع.[27]
- **إزالة هوية مجموعات البيانات الحكومية-التقنيات والحوكمة (NIST SP 800-188):** يقدّم الإطار إرشادات عملية للوكالات الحكومية الأمريكية حول التعامل مع البيانات عند نشرها، مع التركيز على تقنيات إزالة التعرف، ويطرح البيانات المصطنعة كإحدى طرق مشاركة البيانات مع الحفاظ على الخصوصية، كما يناقش التوازن بين منفعة البيانات وبين مخاطر إعادة التعرف، ويقدّم معايير عملية لتقييم هذا التوازن.[28]
- **إرشادات لتقييم ضمانات الخصوصية التباينية (NIST SP 800-226):** هي إطار حديث صدر في عام 2025م، يقدّم إرشادات عملية لتقييم الخصوصية التباينية، بما في ذلك مقاييس كمية لقياس مستوى فقدان الخصوصية. ويُعتبر هذا الإطار ذا صلة وثيقة بالبيانات المصطنعة، بما يضمن التوازن بين الخصوصية والمنفعة.[29]

على الصعيد العملي، برزت عدة مشاريع قدمت أمثلة عملية على الحوكمة في استخدام البيانات المصطنعة، ومن أبرزها تبنّي مكتب التعداد الأمريكي بيانات مُصطنعة لحماية خصوصية أصحاب العمل والموظفين في إحصاءات تنقّل القوى العاملة، كما

أدخل المكتب الخصوصية التباينية كضمان للخصوصية عند نشر نتائج التعداد، وتخضع هذه البيانات لمراجعة دورية للتأكد من أن استخدامها لا يعرّض خصوصية الأفراد للخطر.[30]

الصين

لا تمتلك الصين حتى الآن تشريعاً مستقلاً ينظّم البيانات المصطنعة بوصفها فئة قانونية قائمة بذاتها، غير أنّ هذه البيانات تخضع لأطر تنظيمية أوسع ترتبط بحوكمة البيانات والمحتوى المولَّد بالذكاء الاصطناعي. ويرتكز التنظيم الصيني أساساً على قانون حماية البيانات الشخصية، الذي ينطبق على البيانات القابلة لتحديد هوية الأفراد، بينما تُستثنى البيانات المجهولة بالكامل من نطاقه، ما يعني أن البيانات المصطنعة التي لا تسمح عملياً بإعادة التعرف تُعامل من حيث المبدأ كبيانات خارجة عن نطاق القانون.

على الرغم من ذلك، كانت الصين من أوائل الدول التي وضعت تشريعات تستهدف المحتوى المولد بالذكاء الاصطناعي وخصوصاً التزييف العميق، ففي يناير 2023م، أصدرت الصين "لوائح إدارة خدمات معلومات الإنترنت للتزييف العميق"[31]، التي تركز على منع إساءة استخدام هذا المحتوى في الاحتيال أو نشر المعلومات المضللة، من خلال فرض تسمية واضحة للمحتوى الاصطناعي، إلى جانب إدراج تسمية ضمنية في البيانات الوصفية لضمان تتبّع مصدر المحتوى.

وفي أغسطس 2023م، دخلت "التدابير المؤقتة لإدارة خدمات الذكاء الاصطناعي التوليدي"[32] حيز التنفيذ، والتي وسعت نطاق التنظيم ليشمل دورة تطوير أنظمة الذكاء الاصطناعي نفسها. فقد فرضت هذه التدابير متطلبات تتعلق بدقة وموثوقية وقابلية تتبّع بيانات التدريب، وحماية الملكية الفكرية، ومنع التحيّز والتمييز في البيانات والمخرجات، إلى جانب تعزيز الشفافية والمساءلة. كما نصّت على إلزامية وسم المحتوى المولَّد بالذكاء الاصطناعي ابتداءً من سبتمبر 2025م، وتتكامل هذه المتطلبات مع قوانين أخرى قائمة، مثل قانون أمن البيانات وقانون حماية البيانات الشخصية، لتشكّل إطاراً شاملاً لحوكمة البيانات في نماذج الذكاء الاصطناعي.

ورغم أن هذه الأطر لا تُعرّف البيانات المصطنعة بشكل محدد، إذ إنها تركز على المحتوى المولد بواسطة الذكاء الاصطناعي بشكل عام بهدف مكافحة التزييف العميق، إلا أن المبادئ التي تقوم عليها، ولا سيما الشفافية، وجودة البيانات، وقابلية التتبع، ومنع التحيّز، يمكن أن تنطبق عملياً على البيانات المصطنعة.

سنغافورة

تعتمد سنغافورة استراتيجية تقوم على الإرشادات التقنية والأدوات التشغيلية في حوكمة البيانات المصطنعة. وعلى الرغم من أن قانون حماية البيانات الشخصية يقتصر على تنظيم البيانات الشخصية الحقيقية، إلا أن لجنة حماية البيانات الشخصية أصدرت إرشادات متخصصة جعلت من سنغافورة نموذجاً متقدماً في تنظيم استخدام البيانات المصطنعة من الناحية العملية.

في يوليو 2024م، أصدرت اللجنة الدليل المقترح لتوليد البيانات المصطنعة[5]، ضمن إطار بيئات الاختبار التنظيمية لتقنيات تعزيز الخصوصية، يهدف الدليل إلى توجيه المؤسسات لفهم وتبني البيانات المصطنعة كأداة لحماية الخصوصية ودعم الابتكار في بيئات مُراقبة ومنخفضة المخاطر. ويقدم الدليل إطاراً عملياً متكاملاً لإدارة دورة حياة البيانات المصطنعة عبر خمس مراحل رئيسية وهي: تحديد الأهداف وحالات الاستخدام، وتجهيز البيانات، واختيار منهجيات التوليد وتقييم المنفعة، وتقييم مخاطر إعادة التعرف،

ثم إدارة المخاطر باستخدام ضوابط تقنية وتنظيمية وتعاقدية. كما يركّز الدليل على التوثيق وقابلية التدقيق، من خلال قوائم تحقق تشمل منهجيات التوليد، والخوارزميات المستخدمة، ومعايير الجودة، ويوصي بإعداد قاموس بيانات يوضّح العلاقة بين البيانات المصطنعة والبيانات الحقيقية، ويؤكد الدليل على أن البيانات المصطنعة ليست خالية من المخاطر بطبيعتها، وأن موثوقيتها تعتمد على التقييم المستمر والإدارة الفعّالة للمخاطر.

وبذلك تقدّم سنغافورة نموذجاً متقدماً في الحوكمة التشغيلية للبيانات المصطنعة، يجمع بين المرونة التنظيمية والانضباط التقني ويُنظر إليه بوصفه تجربة قابلة للاقتباس من قبل جهات تنظيمية أخرى تسعى إلى تنظيم هذا المجال دون إعاقة الابتكار.

كندا

يستند النظام الكندي إلى قانون حماية المعلومات الشخصية، ويُتوقع أن يحل محله قريباً قانون حماية خصوصية المستهلك، وكلاهما يعاملان البيانات المجهولة الهوية باعتبارها خارج نطاق تطبيق القانون، بشرط ألا تكون هناك إمكانية للتعرّف على الأفراد، على نحو مشابه لما هو معمول به في اللائحة العامة لحماية البيانات الأوروبية.

ورغم أن التشريعات الحالية لا تتضمن البيانات المصطنعة، إلا أن هناك اهتماماً متزايداً بها، إذ اعتبرها مكتب مفوض الخصوصية الكندي تطوراً نوعياً لتقنيات إزالة الهوية التقليدية، لكنه في الوقت ذاته حذّر من مخاطر احتمالات إعادة التعرف غير المقصودة والتحيزات الإحصائية في البيانات المولدة[33]. كما موّل المكتب دراسة متخصصة حلّلت فوائد ومخاطر الخصوصية المرتبطة بالبيانات المصطنعة في السياق الكندي[34]، وقدمت توصيات عملية لضمان توليد البيانات المصطنعة بشكل آمن، من أبرزها: إزالة جميع المعرفات المباشرة قبل التوليد، واستخدام تقنيات الخلط (Perturbation) المناسبة، وإجراء تقييمات دورية للمخاطر، وتوثيق عملية التوليد بشكل كامل، إلى جانب الالتزام بضوابط تعاقدية وتنظيمية تمنع محاولات إعادة التعرف من قبل الأطراف المستخدمة للبيانات، كما شددت على أهمية أن تحتفظ الجهات المنتجة للبيانات المصطنعة بسجلات تدقيق وتوثيق توضّح مدى قابلية إعادة التعرف، ومدى التزامها بالمعايير التقنية والتنظيمية ذات الصلة.

على مستوى المقاطعات الكندية، بدأت هيئات حماية الخصوصية في مقاطعات كيبيك، وكولومبيا البريطانية في دراسة استخدام البيانات المصطنعة ضمن بيئات اختبار تنظيمية للابتكار في البيانات (Data Innovation Sandboxes)، وهي مبادرة تهدف إلى اختبار حلول مبتكرة تسمح بمشاركة البيانات لأغراض البحث والتطوير مع الحفاظ على الخصوصية[34]. وبذلك يمكن القول إن كندا تبدي اهتماماً بالبيانات المصطنعة وتنظر إليها كأداة امتثال تقنية ضمن منظومة الخصوصية، لا كفئة قانونية مستقلة.

يوضح **الشكل (7)** المشهد التنظيمي لحوكمة البيانات الحقيقية والمصطنعة حول العالم، والذي يُظهر أن الغالبية العظمى من الدول ترتكز على قوانين حماية البيانات الشخصية الحقيقية كإطار تنظيمي رئيسي لحوكمة البيانات، ولا تتضمن هذه الأطر في معظمها تشريعات مخصّصة للبيانات المصطنعة، إلا أن عدداً محدوداً من الدول بدأ بالاعتراف بها ضمن هذه التشريعات، كما في قوانين بعض الولايات الأمريكية والاتحاد الأوروبي، وتضمينها في الغالب ضمن أدوات تعزيز الخصوصية، وتركز بعض التنظيمات على مسألة الشفافية وتسمية المحتوى المصطنع كما في الصين، بينما تبنّت عدد من الدول إرشادات تنظيمية وأدلة تشغيلية، إلى جانب بيئات الاختبار التنظيمية كما في المملكة المتحدة وسنغافورة، ويعكس هذا التباين اختلاف الفلسفات التنظيمية، بين أنظمة تركّز على حماية الخصوصية، وأخرى تركز على الشفافية في المخرجات، وأخرى تعتمد الاتجاه العملي التجريبي. يساعد هذا المشهد التنظيمي غير المتجانس على فهم الفرص المتاحة والتحديات القانونية والتنظيمية التي تواجه الشركات والمطوّرين وصنّاع السياسات، لا سيما فيما يتعلق بتقييم الخصوصية، ومنع إعادة التعرف، وتحقيق التوازن بين المنفعة والامتثال.

الشكل (7): المشهد التنظيمي لحوكمة البيانات الحقيقية والمصطنعة حول العالم. المصدر: مقتبس بتصرف من (Unlocking AI Opportunities with the Responsible Generation and Use of Synthetic Data) آي بي إم.(IBM)[35]

كما يلخص **الجدول (4)** أبرز ملامح السياسات الوطنية تجاه البيانات المصطنعة، والأساس التنظيمي الذي ترتكز عليه، وكيف تنظر إلى البيانات المصطنعة.

الجدول (4): أبرز ملامح السياسات الوطنية تجاه البيانات المصطنعة.

الدولة/ الإقليم	الأساس القانوني والتنظيمي	الموقف من البيانات المصطنعة
الاتحاد الأوروبي	- اللائحة العامة لحماية البيانات. - قانون الذكاء الاصطناعي. - قانون حوكمة البيانات.	يعترف بالبيانات المصطنعة، ولكن دون أحكام مخصصة أو تعريف قانوني محدد، بل يعتبرها أداة امتثال لقوانين الاتحاد وطريقة لحماية الخصوصية.
المملكة المتحدة	- قانون حماية البيانات. - اللائحة العامة لحماية البيانات البريطانية. - إرشادات مكتب مفوض المعلومات.	تُنظم البيانات المصطنعة عبر الأدلة الإرشادية العملية وبيئات الاختبار التنظيمية، وتعتبرها أداة للامتثال والابتكار وحماية الخصوصية.
الولايات المتحدة	- القوانين على مستوى الولايات. - الأطر القطاعية. - معايير وأطر المعهد الوطني للمعايير والتقنية.	تعتمد على تشريعات الولايات والأطر القطاعية التي تُعامل البيانات المصطنعة بوصفها بيانات "مُزالة الهوية". على المستوى الفيدرالي تعتمد على المعايير التقنية الطوعية، دون سنّ تشريع مخصص.
سنغافورة	- قانون حماية البيانات الشخصية. - إرشادات لجنة حماية البيانات الشخصية.	تنظم البيانات المصطنعة عبر الأدلة الإرشادية العملية وبيئات الاختبار التنظيمية، وتعتبرها ضمن تقنيات تعزيز الخصوصية.
الصين	-قانون حماية البيانات الشخصية. - لوائح إدارة خدمات معلومات الإنترنت للتزييف العميق. - التدابير المؤقتة لإدارة خدمات الذكاء الاصطناعي التوليدي.	لا تعرف البيانات المصطنعة بشكل محدد، إذ إنها تركز على الشفافية والإفصاح عن المحتوى المولد بواسطة الذكاء الاصطناعي بشكل عام وخصوصاً مكافحة التزييف العميق.
كندا	- قانون حماية الخصوصية.	تعتبرها تقنية لإزالة الهوية وتحسين الخصوصية، ويُنظر إليها كأداة تقنية لتعزيز الامتثال، وتُختبر تطبيقاتها ضمن بيئات الاختبار التنظيمية.

تقوم المنظمات الدولية بدور مهم في حوكمة التقنيات الناشئة، إذ تتولى وضع المبادئ التوجيهية ورسم المعايير الأخلاقية وتيسير الحوار بين الدول. وفي مجال البيانات المصطنعة، يُعد تدخل هذه المنظمات ضرورياً لإرساء أطر موحدة يمكن للدول والشركات الاسترشاد بها، ومن أهم المنظمات التي عملت على ذلك في سياق البيانات المصطنعة:

الأمم المتحدة

تتبع الأمم المتحدة استراتيجية مزدوجة المسار تجمع بين وضع المبادئ عالية المستوى، وتوفير إرشادات تقنية تفصيلية للجهات التنفيذية. فقد نشرت جامعة الأمم المتحدة (UNU) تقريراً بعنوان "توصيات بشأن استخدام البيانات المصطنعة لتدريب نماذج الذكاء الاصطناعي" [16]يقر بأن هذه التقنية تمثل ممارسة ناشئة ذات إمكانات كبيرة لمعالجة ندرة البيانات، والتحيز، وحماية الخصوصية. ومن الأمثلة التي يوردها التقرير استخدام البيانات المصطنعة لتعزيز مجموعات البيانات الطبية النادرة لتحسين تشخيص الأمراض، أو لنمذجة السيناريوهات الاقتصادية دون الكشف عن معلومات حساسة، أو لمحاكاة الظواهر المناخية لتحسين النماذج التنبؤية. في المقابل يحذر التقرير من مخاطر الأمن السيبراني، والتحيز، مؤكداً على أن الفوائد لا تلغي الحاجة إلى رقابة صارمة تستند إلى معايير أخلاقية واضحة، وشفافية في استخدام البيانات المصطنعة، وتعزيز التعاون الدولي لتطوير المعايير وأفضل الممارسات.

المسار الثاني يتمثل في عمل لجنة الأمم المتحدة الاقتصادية لأوروبا (UNECE)، التي أصدرت دليلاً شاملاً بعنوان "البيانات المصطنعة للإحصاءات الرسمية"[6] مُوجَّهاً إلى المكاتب الإحصائية الوطنية. ويهدف الدليل إلى الموازنة بين منفعة البيانات وحماية الخصوصية، ويقترح أساليب توليد مختلفة بحسب حالة الاستخدام، مع الحفاظ على الخصائص الإحصائية للبيانات الحقيقية، وتكشف هذه الأدلة عن تقسيم واضح للأدوار: فالمبادئ الأخلاقية العامة المتعلقة باستخدام البيانات المصطنعة تضعها جامعة الأمم المتحدة، بينما تتولى وكالات متخصصة مثل (UNECE) صياغة الأدلة التقنية والتشغيلية الموجَّهة للتطبيق العملي.

كما أصدر معهد الأمم المتحدة لبحوث نزع السلاح (UNIDIR) في عام 2024م تقريراً حول حوكمة البيانات المصطنعة في المجالين العسكري والأمني[13]، سلّط فيه الضوء على طبيعة الحوكمة الراهنة التي وصفها بأنها غير مكتملة ومجزأة بين القطاعين المدني والعسكري. وحذّر من مخاطر أمنية دولية تتعلق بإساءة الاستخدام، وتحديات إعادة التعرف، ومعضلات المساءلة. ويرى التقرير في هذه التقنية فرصة لتعزيز التعاون الدولي وتبادل الخبرات دون المساس بالبيانات الحساسة، شرط تطوير آليات تتبع ومعايير موحدة للتوثيق والتسمية، مع وضع ضوابط صارمة للتعامل مع البيانات العابرة للحدود.

تظهر هذه المبادرات أن الأمم المتحدة تتعامل مع البيانات المصطنعة باعتبارها قضية ناشئة في مجال حوكمة البيانات والذكاء الاصطناعي، كما تنظر لها كفرصة لتعزيز الابتكار والخصوصية في آن واحد، لكنها في الوقت نفسه تحذّر من المخاطر المحتملة، وتدعو إلى مزيد من الشفافية، وتوحيد المقاييس، ووضع ضوابط أخلاقية، وتعزيز التنسيق الدولي قبل أن تصبح هذه البيانات واسعة الانتشار بشكل يصعب ضبطه.

منظمة التعاون الاقتصادي والتنمية (OECD)

تؤدي المنظمة دوراً تنسيقياً مهماً في تشكيل حوكمة البيانات المصطنعة، وذلك من خلال عملها الأوسع في مجالات حوكمة البيانات، وحماية الخصوصية، والذكاء الاصطناعي. وبدلاً من سنّ تشريعات ملزمة، تركز المنظمة على تطوير مبادئ توجيهية، وإرشادات، وأطر تحليلية تُستخدم كمرجع للحكومات وصنّاع السياسات. وعبر عدد من إصداراتها، تنظر المنظمة إلى البيانات المصطنعة بوصفها إحدى تقنيات تعزيز الخصوصية القادرة على تمكين مشاركة البيانات، ودعم الابتكار، وتعزيز التعاون عبر الحدود، مع الحد من المخاطر المرتبطة باستخدام البيانات الشخصية[36]. وتؤكد المنظمة -على وجه الخصوص- على دور البيانات المصطنعة

كأداة لتجاوز العوائق القانونية والعملية التي تعيق الوصول إلى البيانات في تطوير أنظمة الذكاء الاصطناعي، ولا سيما في القطاعات شديدة التنظيم أو عالية الحساسية، شريطة أن يُصاحب استخدامها ضوابط صارمة وآليات حوكمة واضحة.[37]

ومن خلال أدوات مثل مبادئ الذكاء الاصطناعي، تشجع المنظمة الحكومات والمؤسسات على تبنّي أطر حوكمة توازن بين منفعة البيانات من جهة، وحماية الخصوصية، والعدالة، والمتانة، والمساءلة من جهة أخرى. وتدعم هذه الأدوات ضمنياً الاستخدام المسؤول للبيانات المصطنعة عبر تعزيز مبادئ الشفافية، وقابلية التفسير، والتصميم المرتكز على الإنسان، وبهذا، ترسخ المنظمة البيانات المصطنعة كأداة حوكمة مشروعة وموثوقة ضمن المنظومة العالمية لحوكمة البيانات والذكاء الاصطناعي، مدمجة ضمن الأطر القائمة لحماية الخصوصية وحوكمة الذكاء الاصطناعي، وليس باعتبارها موضوعاً تنظيمياً منفصلاً.

معهد مهندسي الكهرباء والإلكترونيات (IEEE)

يبرز دور المعهد بوصفه حلقة وصل بين المبادئ الأخلاقية العامة، وصياغة السياسات العليا، والتطبيقات الهندسية العملية. ويستند عمله إلى إطار "التصميم المتوافق أخلاقياً" الذي يهدف إلى دمج القيم الإنسانية والشفافية والمساءلة في تصميم النظم التقنية، وتحويل هذه القيم إلى ممارسات قابلة للتنفيذ من خلال معايير تقنية متخصصة، مثل معيار معالجة الاعتبارات الأخلاقية في التصميم (IEEE 7000) والحد من التحيز الخوارزمي (IEEE 7003). وبهذا يربط المعهد بين المبادئ الدولية الصادرة عن منظمات مثل الأمم المتحدة ومنظمة التعاون الاقتصادي والتنمية، وبين التنفيذ العملي داخل الأنظمة التقنية.

وفيما يخص البيانات المصطنعة، أطلقت هيئة المعايير التابعة للمعهد (IEEE SA) مبادرة تهدف إلى وضع أسس معيارية للخصوصية والدقة في البيانات المصطنعة المولّدة بالذكاء الاصطناعي[38]، إلى جانب تطوير أدوات مشتركة لتقييم تقنيات التوليد. وتسعى المبادرة إلى ترسيخ مفهوم "البيانات المصطنعة العادلة"، أي البيانات التي تُنشأ بطريقة تقلل من إعادة إنتاج التحيزات والتمييز الكامن في البيانات الحقيقية، مع ضمان الشفافية والمساءلة. كما تهدف إلى توحيد المصطلحات، وتحديد أفضل الممارسات وحالات الاستخدام عالية المخاطر في القطاعات الحساسة، ودعم الجهات التنظيمية بأدلة تسهم في بناء أطر امتثال واضحة، بما يعزز الاستخدام المسؤول والعادل للبيانات المصطنعة على المستوى الدولي.

المنظمة الدولية للمقاييس (ISO)

تعمل المنظمة الدولية للمقاييس عبر لجنتها المشتركة مع اللجنة الكهروتقنية الدولية (IEC) المعنية بالذكاء الاصطناعي على تطوير تقرير تقني عن "البيانات المصطنعة في سياق الذكاء الاصطناعي" (ISO/IEC AWI TR 42103) [39] يهدف التقرير إلى توفير نظرة شاملة حول البيانات المصطنعة، مع التركيز على التعاريف والمفاهيم الأساسية، وطرق التوليد، وحالات الاستخدام، إلى جانب الاعتبارات المتعلقة بالخصوصية والتحيز والجودة، وبما أنّ التقرير ما زال في مرحلة الإعداد، فإن تفاصيله النهائية لم تُنشر بعد، غير أنّه من المتوقع أن يساهم في توحيد المفاهيم والمصطلحات بما يوفّر لغة مشتركة بين الجهات التنظيمية والأكاديمية والصناعية.

يوضح **الجدول (5)** أهم الجهات الدولية التي تناولت البيانات المصطنعة ضمن تقارير أو إرشادات أو مشاريع معيارية، مبيّناً طبيعة الإصدارات، ونوعها ومدى صلتها المباشرة بتطوير أطر حوكمة البيانات المصطنعة.

الجدول (5): أبرز المبادرات الدولية ذوات الصلة بحوكمة البيانات المصطنعة.

الجهة		الإصدار	النوع	الصلة بحوكمة البيانات المصطنعة
الأمم المتحدة	جامعة الأمم المتحدة (UNU)	توصيات بشأن استخدام البيانات المصطنعة لتدريب نماذج الذكاء الاصطناعي.	إرشادات	تسلط التوصيات الضوء على فرص ومخاطر البيانات المصطنعة وتدعو إلى التعاون الدولي ووضع معايير دولية للشفافية والجودة.
	لجنة الأمم المتحدة الاقتصادية لأوروبا (UNECE)	البيانات المصطنعة للإحصاءات الرسمية.	إرشادات	يوجه الدليل مكاتب الإحصاء الوطنية إلى استخدام البيانات المصطنعة، ويقترح أساليب توليد تتوافق مع متطلبات التوثيق والشفافية واختبارات إعادة التعرف.
	معهد الأمم المتحدة لبحوث نزع السلاح (UNIDIR)	تبعات حوكمة البيانات المصطنعة في سياق الأمن الدولي.	تقرير	يناقش التطبيقات العسكرية للبيانات المصطنعة، ويحذّر من المخاطر المحتملة لها على الأمن الدولي، ويدعو إلى معايير دولية موحّدة.
هيئة المعايير في معهد مهندسي الكهرباء والإلكترونيات (IEEE SA)		البيانات المصطنعة العادلة.	تقرير (قيد الإعداد)	يهدف إلى تعزيز العدالة والجودة والشفافية في استخدام البيانات المصطنعة عالمياً.
المنظمة الدولية للمقاييس (ISO) واللجنة الكهروتقنية الدولية (IEC)		البيانات المصطنعة في سياق الذكاء الاصطناعي (ISO/IEC AWI TR).	تقرير (قيد الإعداد)	يقدم نظرة على المفاهيم الأساسية للبيانات المصطنعة وطرق التوليد والاستخدامات. ومن المتوقع أن يساهم في توحيد المفاهيم وتمهيد معايير مستقبلية للخصوصية والجودة.

أبرز الاتجاهات

يظهر تحليل المشهد التنظيمي للبيانات المصطنعة أن الإطار التشريعي الخاص بها لا يزال ناشئاً، ويعرض **الجدول (6)** المحاور الأساسية لحوكمة البيانات المصطنعة، موضحاً لكل محور الآليات التنظيمية المطبَّقة وأمثلة عملية من القوانين والإرشادات والأطر الدولية ذات الصلة، بما يسهِّل على صُنّاع السياسات فهم ملامح بيئة الحوكمة عالمياً.

الجدول 6: محاور حوكمة البيانات المصطنعة وأبرز آليات تنفيذها وأمثلة دولية لتطبيقها.

المحور	الآلية	أمثلة
حماية الخصوصية	تقنيات تعزيز خصوصية البيانات.	**الولايات المتحدة الأمريكية:** "إرشادات لتقييم ضمانات الخصوصية التفاضلية" NIST SP 800-226) **المملكة المتحدة:** حزمة الإرشادات لتعزيز الخصوصية التي قدمها مكتب مفوض المعلومات، وتتضمن تقنيات مختلفة لتعزيز الخصوصية مثل الخصوصية التفاضلية.
	اختبارات تقييم خطر إعادة التعرف.	**المملكة المتحدة:** حزمة الإرشادات لتعزيز الخصوصية التي قدمها مكتب مفوض المعلومات، وتتضمن اختبارات عملية مثل إمكانية الفصل، والربط، والاستدلال، واختبار المتسلل المحفَّز. **سنغافورة:** قدّم الدليل المقترح لتوليد البيانات المصطنعة"، ضمن أفضل الممارسات لتعزيز الخصوصية إجراء اختبارات لتقييم خطر إعادة التعرف.
	منع أو تجريم إعادة التعرف.	**المملكة المتحدة:** قانون حماية البيانات المادة 171) تجعل إعادة تحديد هوية بيانات أزيلت هويتها دون إذن جريمة. كما يوصي تقرير هيئة السلوك المالي بإدراج التزامات تعاقدية واضحة عند مشاركة البيانات المصطنعة مع أطراف خارجية. **سنغافورة:** قدّم الدليل المقترح لتوليد البيانات المصطنعة"، ضمن أفضل الممارسات فرض ضوابط تعاقدية تمنع محاولات إعادة التعرف. **الولايات المتحدة الأمريكية:** تشترط تشريعات الخصوصية في بعض الولايات التعهّد العلني والإلزام التعاقدي بعدم إعادة التعرف.
الجودة والمنفعة	موازنة المنفعة والخصوصية.	**المملكة المتحدة:** قدمت هيئة السلوك المالي إطاراً يوازن بين الخصوصية والمنفعة والمصداقية. **سنغافورة:** قدم "الدليل المقترح لتوليد البيانات المصطنعة"، إطاراً يضم اختبارات للمنفعة مقابل مخاطر إعادة التعرّف.
التحيز وتحسين التمثيل	تقليل التحيز وكشفه.	**الاتحاد الأوروبي:** يجيز قانون الذكاء الاصطناعي المادة 10 (5)(أ) استخدام البيانات المصطنعة لتصحيح التحيّز، أو عندما تكون البيانات الحقيقية غير كافية. **سنغافورة:** قدّم الدليل المقترح لتوليد البيانات المصطنعة" إرشادات عملية لرصد التحيز ومعالجة عدم توازن البيانات لضمان العدالة.
الشفافية	الإفصاح عن المحتوى المولد باستخدام نماذج ذكاء اصطناعي (التزييف العميق)	**الولايات المتحدة الأمريكية:** إرشادات NIST AI 100-4) حول توثيق وتسمية المحتوى، وتتبع المصدر والكشف عن المحتوى المصطنع. **الصين:** لوائح التوليف العميق" و"التدابير المؤقتة للذكاء الاصطناعي التوليدي"، التي تتطلب تسمية صريحة وضمنية للبيانات المولدة بالذكاء الاصطناعي، كما تلزم مزوّدي الخدمات بالحفاظ على سجلات التوليد والتعديلات اللاحقة لأغراض التدقيق والرقابة. **الاتحاد الأوروبي:** قانون الذكاء الاصطناعي المادة 50 (2) تفرض متطلبات تتعلق بالشفافية، وذلك بإلزام أنظمة الذكاء الاصطناعي التي تولد محتوى مصطنعاً بوضع علامات لتسمية هذه المخرجات على أنها مولدة اصطناعياً.
	الإفصاح عن أن البيانات مصطنعة، وتوثيق منهجية توليد البيانات.	**الولايات المتحدة الأمريكية:** إرشادات NIST AI 100-4) توصي بتوثيق مصدر وطريقة توليد البيانات والاحتفاظ بسجلات قابلة للتدقيق للبيانات المصطنعة. **الاتحاد الأوروبي:** قانون الذكاء الاصطناعي المواد 11–13) تتناول الشفافية وتوثيق بيانات التدريب، وتنص على وجوب توثيق مصدر البيانات. **المملكة المتحدة:** إرشادات مكتب مفوض المعلومات التي توصي بالتوثيق في سجل المعالجة إذا كانت البيانات مشتقة من بيانات شخصية في الأصل، والإفصاح للمستخدمين والجهات الرقابية عن عمليات توليد بيانات مصطنعة لأغراض التحقق. **سنغافورة:** قدم "الدليل المقترح لتوليد البيانات المصطنعة" متطلبات توثيق تفصيلية. تتضمن الإفصاح داخلياً عن استخدام بيانات مصطنعة ضمن النظم المؤسسية، وعند مشاركة البيانات خارجياً (مع باحثين أو أطراف ثالثة)، يجب أن تتضمن الوثائق المرافقة إشارة واضحة إلى أن البيانات مصطنعة بالكامل أو جزئياً، مع توثيق منهجية التوليد. **الصين:** لوائح "التوليف العميق" و"التدابير المؤقتة للذكاء الاصطناعي التوليدي" تلزم مزوّدي الخدمات بالحفاظ على سجلات التوليد لأغراض التدقيق والرقابة.

الفجوات التنظيمية

يتضح من المراجعة السابقة أن حوكمة البيانات المصطنعة ما زالت في طور التبلور عالمياً، مع بروز عدد من الفجوات في الجهود التنظيمية، من أبرزها:

غياب تعريف موحّد

لا يزال الوضع القانوني للبيانات المصطنعة غير موحد عالمياً، في ظل غياب توافق دولي حول تعريف هذه البيانات، ففي الولايات المتحدة بدأت بعض الولايات بوضع تعريف لها، مثل قانون خصوصية المستهلك في ولاية يوتا، الذي عرّفها بأنها بيانات مولَّدة بخوارزميات أو نماذج إحصائية ولا تحتوي على معلومات شخصية، في المقابل، لم تعتمد الأطر القانونية الكبرى في الاتحاد الأوروبي تعريفاً مباشراً لها، بل تدرجها ضمن تقنيات تعزيز الخصوصية، بينما لا تُذكَر أبداً في عدد من القوانين، هذا التباين يكشف عن فجوة تنظيمية واضحة ويؤكد على الحاجة إلى وضوح وتوافق أكبر، وتستلزم الحوكمة صياغة تعريف قانوني دقيق يحدد ماهية البيانات المصطنعة، باختلاف أنواعها المصطنعة كلياً، أو جزئياً، أو الهجينة.

إشكالية تصنيف البيانات المصطنعة ضمن نطاق قوانين حماية البيانات

تبرز مسألة التصنيف القانوني للبيانات المصطنعة كأحد أبرز الفجوات التنظيمية، خاصة إذا كانت هذه البيانات مولّدة من بيانات شخصية، إذ لا يزال من غير الواضح ما إذا كانت البيانات المصطنعة تعد بيانات شخصية تنطبق عليها قوانين حماية البيانات الشخصية، أم مجهولة تماماً، تنطلق بعض التفسيرات للقانون من افتراض مبدئي مفاده أنه إذا كانت البيانات الحقيقية بيانات شخصية، فإن البيانات المصطنعة الناتجة عنها تُعد كذلك بيانات شخصية، إلا إذا أُثبت بدرجة عالية من الثقة أن خطر إعادة التعرف على الأفراد ضئيل جداً، ولكن هذا يتطلب تقييماً صارماً لمخاطر إعادة التعرف وهو ما يعد تحدياً خصوصاً في ظل غياب معايير أو قيم كمية تحدد إمكانية تحديد هوية الأفراد من البيانات المصطنعة. لذلك تستلزم الحوكمة تحديد معايير تحدد متى تكون البيانات المصطنعة شخصية ومتى تكون مجهولة تماماً، وما يترتب على ذلك من خضوعها لقوانين حماية البيانات الشخصية القائمة أو خروجها عنها.

مدى ملاءمة الأطر القانونية الحالية للبيانات المصطنعة

رغم التطور السريع في تقنيات توليد البيانات المصطنعة، ما زالت هناك حاجة إلى تحديد كيفية تطبيق القوانين الحالية الخاصة بحماية البيانات والخصوصية على البيانات المصطنعة، وكيف يمكن تكييفها لسد الثغرات التنظيمية. إذ لا يوجد حتى الآن إجماع بين الخبراء حول ما إذا كانت القوانين القائمة كافية، أم أن هناك حاجة إلى تشريعات أو أطر متخصصة للبيانات المصطنعة، فالتشريعات الحالية تركّز على حماية البيانات الحقيقية، ولم تصمم للتحكم في نماذج الذكاء الاصطناعي التوليدي والبيانات التي تنتج عنها. خصوصاً وأن ذلك قد يتطلب أن تنتقل الحوكمة من التركيز على حماية البيانات إلى تنظيم آليات توليد البيانات ذاتها[15]. كما تثير النقاشات القانونية تساؤلات حول قدرة أطر حماية البيانات القائمة على معالجة الإشكالات الفريدة للبيانات المصطنعة. فعلى سبيل المثال، لم تتناول اللوائح القائمة بعد مسألة التطابق العرضي (Coincidental Matching)، وهي الحالة التي قد تتطابق فيها بيانات مصطنعة مع أفراد حقيقيين بمحض الصدفة، مما يخلق خطراً غير مقصود يمس الخصوصية.[7]

يستدعي تقييم هذه المخاطر دراسة الدور الحقيقي لقوانين حماية البيانات ضمن المنظومة القانونية الأوسع، وتحديد نطاق الأضرار التي تسعى إلى منعها. فالتوسّع في تطبيق قوانين حماية البيانات على جميع أشكال البيانات المصطنعة دون مراعاة لطبيعتها التقنية، قد يؤدي إلى الانحراف عن الغاية الأصلية لهذه القوانين التي وُضعت أساساً لحماية بيانات الأفراد من المعالجة غير المشروعة، لا لتنظيم التوليد الإحصائي أو الخوارزمي للبيانات. وبناءً على ذلك، فإن التحدي أمام الحوكمة لا يتمثل في إنشاء تشريعات جديدة فحسب، بل في إعادة تفسير الأطر الحالية بما يضمن حماية الخصوصية دون تقييد الابتكار القائم على البيانات المصطنعة، وتحقيق توازن مستدام بين السلامة والمنفعة.

استخدام البيانات الشخصية لتوليد بيانات مصطنعة

يُثير توليد البيانات المصطنعة من بيانات شخصية مسألة قانونية عندما تولد من بيانات شخصية، والذي يُعد في عدد من الأطر التنظيمية، شكلاً من أشكال معالجة البيانات الشخصية، ما قد يُخضع الجهات المنتجة لالتزامات قانونية، فوفقاً للائحة العامة لحماية البيانات[1]، إذا استلزم إنشاء البيانات المصطنعة معالجة بيانات شخصية، فإن ذلك يخضع لقواعد تنظيمية. وهذا يعني أن الجهة المنتِجة مُلزَمة بالامتثال لمبادئ أساسية، من أبرزها، مبدأ تقليل البيانات، أي الاكتفاء باستخدام أقل قدر ممكن من البيانات الضرورية لهذا الغرض، مع إزالة أو تشفير أي معرفات مباشرة للأفراد قبل التوليد، كذلك تستلزم ضمان أمن البيانات الحقيقية أثناء عملية التوليد، بما يشمل منع أي وصول غير مصرح به، وكذلك وجود أساس قانوني مشروع لمعالجة البيانات الشخصية بغرض إنشاء بيانات مصطنعة، مثل الموافقة، أو المصلحة المشروعة، أو غرض البحث العلمي.

المسؤولية القانونية

تمثل المسؤولية القانونية إحدى الركائز المهمة في مناقشات حوكمة البيانات المصطنعة، إذ يتعيّن تحديد من يتحمل المسؤولية عند وقوع خطأ نتيجة استخدام هذه البيانات. ويمكن حدوث ذلك في حالات متعددة مثل إنتاج بيانات قابلة لإعادة التعرف وكشف هوية أشخاص، أو تشخيص طبي غير دقيق بناء على هذه البيانات، وفي هذه الحالات، يصعب تحديد المسؤول المباشر ما يطرح الحاجة إلى أطر قانونية واضحة لتحديد المسؤوليات. لا تُعالج القوانين الحالية مسألة المسؤولية عن البيانات المصطنعة، فهي تركز بالأساس على البيانات الشخصية، هذا الفراغ التشريعي قد يفتح الباب أمام تحديث القوانين كي تغطي البيانات المصطنعة، مع استحداث آليات مثل المساءلة المهنية التي تلزم المؤسسات بتعيين مسؤول محدّد يوافق رسمياً على إصدار البيانات المصطنعة، أو المسؤولية الصارمة في التطبيقات عالية المخاطر، والتي يتحمل فيها مولّد البيانات المصطنعة المسؤولية عن الأضرار اللاحقة بغض النظر عن مستوى العناية التي بذلها.

غياب معايير دولية للبيانات المصطنعة

تُعدّ المعايير الدولية الموحدة أداةً مهمةً للحوكمة، ورغم عدم وجود معايير دولية خاصة بالبيانات المصطنعة على المستوى الدولي حتى الآن، إلا أن العمل جارٍ على تطويرها، إذ أطلق معهد مهندسي الكهرباء والإلكترونيات (IEEE) مبادرة خاصة يتوقع أن تشمل جوانب مثل مقاييس الجودة، وضمانات الخصوصية، والتوثيق، وقابلية التشغيل البيني، وتوفير أدوات تقييم مشتركة للتقنيات المستخدمة في توليد البيانات، كما أنها ستساعد على مشاركة البيانات المصطنعة بين الدول. وبالتوازي، تعمل المنظمة الدولية للمقاييس (ISO) بالتعاون مع اللجنة الكهروتقنية الدولية (IEC) على إعداد تقرير تقني حول البيانات المصطنعة لتوحيد المفاهيم وتوضيح أساليب التوليد وحالات الاستخدام، تمهيداً لوضع معايير دولية أكثر تفصيلاً في المستقبل.

رابعاً: الوضع الراهن في المملكة

تشهد المملكة العربية السعودية تطوراً كبيراً في منظومة البيانات الوطنية، مدفوعاً برؤية السعودية 2030 التي تضع البيانات والذكاء الاصطناعي في صميم التحول الرقمي وصنع القرار. وعلى الرغم من أن البيانات المصطنعة لا تزال في مرحلة التشكّل من حيث التنظيم الصريح، فإن الأطر التشريعية والمؤسسية القائمة تهيئ بيئة داعمة لاستخدامها بصورة آمنة ومسؤولة. يستعرض هذا الفصل الجهود الوطنية التي تبذلها الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا) في مجال البيانات، ويحلل موقع البيانات المصطنعة ضمن الإطار التنظيمي القائم، مع إبراز المبادرات الحديثة ذات الصلة بتقنيات تعزيز الخصوصية والابتكار المنظم.

جهود المملكة في بناء منظومة البيانات الوطنية

تبذل الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا) جهوداً كبيرة في بناء منظومة وطنية متكاملة لإدارة البيانات وحوكمتها، بما يعزز استخدامها كأصل وطني داعم للتحول الرقمي. وقد شملت هذه الجهود تطوير أطر تنظيمية وتشريعية متينة لتنظيم جمع البيانات ومعالجتها ومشاركتها وحمايتها، وفي مقدمتها نظام حماية البيانات الشخصية، وسياسات البيانات المفتوحة، وتصنيف البيانات، ومشاركتها، بما يضمن حماية الخصوصية، وتعزيز الشفافية، وخلق بيئة آمنة لتبادل البيانات بين الجهات الحكومية. كما دعمت المملكة هذه الأطر بسلسلة من الوثائق التنظيمية واللوائح التنفيذية التي تحدد ضوابط إدارة البيانات وحوكمتها ونقل البيانات الشخصية، بما يسهم في الحد من إساءة استخدام البيانات وضمان الاستخدام المسؤول لتقنيات الذكاء الاصطناعي. وعلى المستوى التشغيلي، أنشأت المملكة بنية تحتية متقدمة تشمل بنك البيانات الوطني، والسحابة الحكومية، ومنصات التحليل والاستشراف، ومنصة حوكمة البيانات الوطنية، بهدف رفع جودة البيانات، وتسهيل مشاركتها، وتمكين التحليل المتقدم وصنع السياسات المبنية على البيانات. تشكّل هذه الجهود أساساً تنظيمياً وتقنياً داعماً متيناً لا سيما في السياقات التي تتطلب تقليل المخاطر المرتبطة بالبيانات الشخصية وتعزيز الابتكار الآمن.

الإطار التنظيمي وعلاقته بالبيانات المصطنعة

ترتكز حماية البيانات في المملكة العربية السعودية على نظام حماية البيانات الشخصية، الذي دخل حيّز النفاذ في سبتمبر 2023م، وتتولى الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا) الإشراف على تنفيذه وضمان الالتزام بأحكامه. ويعرّف نظام البيانات الشخصية بأنه كل بيان —بصرف النظر عن مصدره أو شكله— من شأنه أن يؤدي إلى التعرف على الفرد بصورة مباشرة أو غير مباشرة، كما يُلزم كلاً من المتحكم والمعالج بتطبيق مبادئ الشفافية، والمشروعية، والتقليل من جمع البيانات، وحمايتها عبر تدابير تقنية وتنظيمية مناسبة. وعلى الرغم من أن البيانات المصطنعة لم تُذكر صراحةً في نص النظام ولا في لائحته التنفيذية، فإن اللائحة تضمنت مفهوم "إخفاء الهوية"، وعرّفته بأنه إزالة المعرفات المباشرة وغير المباشرة بشكل يتعذر معه تحديد هوية صاحب البيانات. وتنص المادة (9) من اللائحة التنفيذية على أن البيانات التي أُخفيت هوية أصحابها لا تُعد بيانات شخصية، وبالتالي لا تنطبق عليها أحكام نظام حماية البيانات الشخصية. وبناءً عليه، فإن الوضع القانوني للبيانات المصطنعة في المملكة يرتبط بإمكانية إعادة التعرف، فإذا كانت البيانات تسمح بتحديد هوية الشخص بشكل مباشر أو غير مباشر فإنها تخضع للنظام، أما إذا أُخفيت الهوية فيها بصورة تمنع إعادة التعرف، فإنها تُعد بيانات مجهولة وتخرج من نطاق النظام.

على المستوى الإرشادي والتشغيلي، أصدرت سدايا إرشادات للتزييف العميق تهدف إلى تنظيم التعامل مع الوسائط المصطنعة، بما في ذلك الصور والصوت والفيديو والنصوص، والتي قد تُستخدم في الانتحال أو التضليل أو انتهاك الخصوصية. وتتضمن هذه الإرشادات مبادئ أساسية مثل الشفافية، وتسمية المحتوى المصطنع عند النشر، والمساءلة عن الأضرار الناتجة. ورغم أن هذه الإرشادات تركز على التزييف العميق، فإنها تعكس توجهاً تنظيمياً أوسع نحو حوكمة المخرجات المصطنعة، وهو ما يمكن الاستفادة منه عند تطوير ونشر مجموعات بيانات مصطنعة، خاصة فيما يتعلق بالشفافية، وتخفيف التحيّز، والمسؤولية.

كما أطلقت الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا) أول بيئة تنظيمية تجريبية للبيانات والخصوصية (Regulatory Sandbox) في المنطقة عام 2023م، بما يعزز مكانة المملكة في مجال حماية البيانات والابتكار المنظم. تهدف هذه البيئة إلى توفير مساحة آمنة للابتكار تجمع الخبراء، وصناع السياسات، ورواد الأعمال، لاستكشاف الحلول المتقدمة لتعزيز الخصوصية، وأفضل الممارسات التنظيمية. ويركز البرنامج على دعم الابتكار المسؤول في تطوير تقنيات تعزيز الخصوصية ضمن بيئة تنظيمية موثوقة، بما يفتح المجال أمام اختبار حلول متقدمة، من بينها البيانات المصطنعة، دون الإخلال بمتطلبات حماية البيانات الشخصية. وتشمل أهداف البرنامج إطلاق بيئة اختبار للابتكارات في تقنيات تعزيز الخصوصية، وتعزيز الحوار المؤسسي بين الجهات التنظيمية والمبتكرين.

وعلى الرغم من أن الأطر التنظيمية القائمة توفر أساسًا مهمًا لحوكمة البيانات وحمايتها، فإن الطبيعة المستجدة لبعض استخدامات البيانات المصطنعة قد تستدعي تطوير أدلة إرشادية أو توضيحات تنظيمية متخصصة تحدد المتطلبات والضوابط المناسبة بحسب حالة الاستخدام ومستوى المخاطر.

خامساً: الدروس المستفادة وخارطة الطريق

يعتمد نجاح حوكمة البيانات المصطنعة على فهم التحديات التنظيمية والتقنية والأخلاقية المحيطة بها والاستجابة لها بشكل متوازن. يهدف هذا الفصل إلى استخلاص أبرز الدروس من التجارب الدولية وتقديم خارطة طريق عملية للتبني. بدءاً بتسليط الضوء على الغموض التنظيمي والتحديات التقنية التي تعيق الاستخدام الآمن والفعال للبيانات المصطنعة، إلى إبراز أهمية المعايير الموحدة والخصوصية السياقية وأهمية المنهجية متعددة الأطراف. كما يطرح تسلسلاً من الخطوات العملية، من تحديد الأهداف وتطوير الأطر التنظيمية إلى تطبيق الضوابط التقنية وتعزيز الشراكات، لضمان استخدام البيانات المصطنعة بصورة تخدم المجتمع والاقتصاد.

الدروس المستفادة

1. **الموازنة بين الخصوصية والمنفعة:** يرتبط تأثير البيانات المصطنعة في الخصوصية بشكل مباشر بالسياق وتنوع الاستخدامات، إذ تختلف المتطلبات باختلاف القطاع؛ فالمجالات الحساسة كالرعاية الصحية والقطاع العسكري تحتاج إلى ضوابط رقابية صارمة ومنهجية تنظيمية مخصصة تراعي بدقة تقييم مخاطر إعادة تحديد الهوية وجودة البيانات، في حين تسمح القطاعات الأخرى بمساحة أكبر من المرونة، مما يجعل من الصعب وضع إطار تنظيمي موحد يناسب كافة القطاعات. ويبقى التحدي الأهم في تحقيق التوازن بين حماية الخصوصية وضمان بقاء البيانات مفيدة لأغراض البحث والتدريب. وتعد البيانات المصطنعة إحدى الوسائل الفعالة لحماية الهوية، لكن نجاحها في تحقيق هذا التوازن يتوقف على اختيار خوارزميات التوليد الملائمة، بالإضافة إلى تطبيق إجراءات داعمة مثل إزالة القيم الشاذة، وإضافة التشويش، واستخدام تقنيات التعمية للحد من إمكانية إعادة تحديد هوية الأفراد.
2. **الشراكة متعددة الجهات والإدارة التكيفية:** من أبرز الدروس المستفادة ضرورة إشراك جميع الأطراف المعنية في حوكمة البيانات المصطنعة، إذ يؤكد تقرير الأمم المتحدة على أن الأطر الفعّالة يجب أن تجمع بين الحكومات والمؤسسات الصناعية والأكاديميين والمجتمع المدني لبناء شراكات تحقق الثقة وتعزز المسؤولية في تطوير هذه البيانات. كما أن التطور المتسارع في تقنيات الذكاء الاصطناعي وظهور نماذج توليد جديدة يفرض الحاجة إلى أُطر تنظيمية تتسم بالمرونة وقابلية التكيف مع المخاطر المستجدة. ويستلزم ذلك مراجعة السياسات بانتظام وتحديثها لمواكبة التطورات التقنية، وضمان استجابة تنظيمية استباقية وديناميكية.

خارطة الطريق

لترجمة الدروس المستفادة إلى إجراءات ملموسة، تقترح خارطة الطريق التالية منهجية مرحلية ومنظمة لتبني البيانات المصطنعة، توازن بين الابتكار والمسؤولية.

المرحلة الأولى: التأسيس والاستراتيجية

الخطوة 1: تحديد الأهداف وحالات الاستخدام بوضوح

قبل الشروع في توليد البيانات المصطنعة، يجب على المؤسسات تحديد أهدافها بدقة. فحالة الاستخدام هي التي تحدد التوازن المطلوب بين المنفعة والدقة والخصوصية. يجب على المؤسسات أن تسأل: هل الهدف هو اختبار البرمجيات، أم تدريب نماذج تعلم الآلة، أم مشاركة البيانات للبحث العام؟ الإجابة عن هذا السؤال ستوجه اختيار التقنيات ومعايير التقييم المناسبة، وتضمن أن تكون الجهود المبذولة متناسبة مع الغرض النهائي.

الخطوة 2: بناء إطار حوكمة شامل

يجب إنشاء إطار حوكمة داخلي واضح قبل تبني التقنية على نطاق واسع. يتضمن هذا الإطار تحديد الأدوار والمسؤوليات، ووضع سياسات للموافقة على المشاريع، وإنشاء عمليات لتقييم المخاطر وإدارتها. يجب أن يرتكز هذا الإطار على مبادئ المساءلة، لضمان وجود جهة مسؤولة عن جودة البيانات المصطنعة وسلامتها وآثارها الأخلاقية.

المرحلة الثانية: التطوير والتنفيذ التقني

الخطوة 3: وضع معايير تقنية وبروتوكولات للتحقق

لمواجهة تحدي غياب المعايير الموحدة، يجب على المؤسسات تطوير بروتوكولات داخلية للتحقق من صحة البيانات المصطنعة. يجب أن تتجاوز هذه البروتوكولات المقارنات الإحصائية البسيطة لتشمل مقاييس تقيّم أداء البيانات في المهام المحددة لها. كما أن توثيق عملية التوليد بأكملها، بما في ذلك مصدر البيانات الحقيقية، والنموذج المستخدم، ومعاملات الضبط المطبقة يعد أمراً بالغ الأهمية لتحقيق الشفافية وقابلية التدقيق، وهو ما يتماشى مع أطر إدارة المخاطر المطبقة في إطار المعهد الوطني للمعايير والتقنية (NIST).

الخطوة 4: تطبيق ضوابط الشفافية وتتبع المصدر

لبناء الثقة ومنع إساءة الاستخدام، من الضروري تطبيق آليات لتتبع مصدر المحتوى المصطنع. يمكن أن يشمل ذلك تقنيات مثل العلامات المائية الرقمية والبيانات الوصفية (Metadata) المشفرة. إن تبني متطلبات تسمية واضحة، على غرار قانون الذكاء الاصطناعي في الاتحاد الأوروبي واللوائح الصينية، يضمن إعلام المستخدمين بأنهم يتفاعلون مع محتوى أنشئ بواسطة الذكاء الاصطناعي، مما يعزز الشفافية ويقلل من مخاطر التضليل.

المرحلة الثالثة: بناء المنظومة والتطور المستمر

الخطوة 5: تشجيع الابتكار عبر الشراكات وبيئات الاختبار

تحفيز الابتكار المسؤول من خلال إنشاء بيئات اختبار تنظيمية (Regulatory Sandboxes)، على غرار تلك التي تديرها هيئة السلوك المالي في المملكة المتحدة، والتي تسمح للشركات باختبار حلول البيانات المصطنعة في بيئة خاضعة للرقابة، كما أن تمويل البحث والتطوير الأساسي، كما تفعل مؤسسة العلوم الوطنية الأمريكية (NSF)، يضمن استمرار تطور التقنية لمواجهة التحديات المستقبلية.

الخطوة 6: تبني منهجية تنظيمية مرنة ومستقبلية

نظراً إلى التطور السريع لتقنيات الذكاء الاصطناعي، يجب أن تكون الأطر التنظيمية مرنة وقادرة على التكيف، فبدلاً من وضع قواعد صارمة ومحددة، يجب أن تركز السياسات على المبادئ الأساسية مثل العدالة والشفافية والمساءلة. تتطلب هذه المنهجية مراجعة مستمرة للسياسات لمواكبة التطورات التقنية. ورصد المخاطر الناشئة مثل تدهور النموذج (Model Collapse)، حيث يمكن أن يؤدي التدريب المتكرر على البيانات المصطنعة إلى تدهور جودة النماذج بمرور الوقت.

التوصيات

لتحقيق ذلك، تقترح الدراسة مجموعة من الممارسات المسؤولة الموجَّهة إلى كافة الأطراف الفاعلة.

توصيات للمستفيدين

- **تحديد حالة الاستخدام بوضوح:** قبل الشروع في تبني البيانات المصطنعة، من الضروري تحديد الغرض بدقة؛ هل الهدف اختبار البرمجيات أم تدريب النماذج أم مشاركة البيانات لأغراض بحثية؟ تحديد حالة الاستخدام يُمكّن من تحديد مستوى الدقة والخصوصية المطلوب واختيار الأدوات المناسبة.
- **إجراء تقييمات صارمة للمخاطر:** لا ينبغي الافتراض بأن البيانات المصطنعة توفر خصوصية مطلقة، بل على المؤسسات إجراء تقييمات دقيقة لاحتمالية إعادة التعرف، خاصة عند التعامل مع مجموعات بيانات تحتوي على معلومات حساسة.
- **المطالبة بالشفافية من الموردين:** عند الحصول على بيانات مصطنعة أو أدوات لتوليدها من جهات خارجية، يجب المطالبة بمستندات تفصيلية عن مصادر البيانات الحقيقية، والنموذج المستخدم في التوليد، والتحيزات المعروفة، والقيود المفروضة على النموذج، لضمان المساءلة.
- **الاستثمار في الحوكمة الداخلية وبناء الخبرات:** يتعين إنشاء أطر حوكمة داخلية تحدد الأدوار والمسؤوليات المتعلقة باستخدام البيانات المصطنعة، مع الاستثمار في تدريب الفرق على فهم الجوانب التقنية والأخلاقية والقانونية لهذه التقنية.

توصيات للمطورين

- **دمج الاعتبارات الأخلاقية في التصميم:** ينبغي ألا تكون معالجة التحيز فكرة لاحقة؛ بل يتعين تقييم البيانات الأولية بحثاً عن التحيزات الكامنة وتطبيق تقنيات للتخفيف منها خلال عملية التوليد، بهدف إنشاء بيانات أكثر عدلاً من الواقع نفسه.
- **اعتماد الشفافية والتوثيق الشامل:** يجب توثيق كل خطوة من خطوات عملية التوليد، بدءاً من خصائص البيانات الأولية، مروراً بالنموذج المستخدم ومعاملاته، وصولاً إلى مقاييس التقييم، بما يضمن بناء الثقة وإمكانية التدقيق المستقل.
- **تطوير مقاييس تحقق متقدمة:** لا تقتصر عملية التحقق على مقارنة التوزيعات الإحصائية؛ بل ينبغي تطوير مقاييس تقيس أداء البيانات المصطنعة في مهام واقعية، وتقيّم مدى تمثيلها للحالات النادرة والمتطرفة.
- **دمج التقنيات المعززة للخصوصية:** استخدام تقنيات مثل الخصوصية التباينية لتوفير ضمانات ضد تسرب المعلومات، مع توضيح المقايضات بين مستوى الخصوصية المحقق وفائدة البيانات الناتجة.

الخاتمة

أصبحت البيانات المصطنعة أحد المرتكزات الرئيسية للابتكار في منظومة الذكاء الاصطناعي المعاصرة، لما توفره من إمكانات كبيرة لمعالجة ندرة البيانات، وحماية الخصوصية، وتسريع البحث والتطوير، لكنها في المقابل لا تخلو من تحديات تنظيمية وأخلاقية عميقة، ما يجعل التعامل مع البيانات المصطنعة قضية استراتيجية تتجاوز بعدها التقني إلى أبعاد قانونية ومجتمعية أوسع. وتُظهر الدراسة أن تعظيم الاستفادة من هذه التقنية يتطلب موازنة دقيقة بين تشجيع الابتكار من جهة، ووضع الأطر التي تصون الحقوق والقيم المجتمعية من جهة أخرى.

ويعكس المشهد العالمي غياب إطار دولي موحّد وتوافق عالمي لحوكمة البيانات المصطنعة، ويؤكد على أن مستقبل هذه التقنية يواجه تحديات جوهرية، من بينها الموازنة بين فائدة البيانات ومستوى الخصوصية، واحتمال إعادة إنتاج التحيزات أو تضخيمها، إضافة إلى المخاطر المنهجية المرتبطة بالإفراط في الاعتماد على البيانات المصطنعة، مثل انهيار النماذج أو انحرافها عن السياقات الواقعية.

وتشير الاتجاهات المستقبلية إلى تطور البيانات المصطنعة من مجرد محاكاة للبيانات الموجودة على أرض الواقع إلى بناء عوالم بيانات جديدة بالكامل، مثل ما يُعرف بالبيانات فائقة الاصطناع (Hypersynthetic Data)، حيث تُدمج المحاكاة الفيزيائية والتوائم الرقمية والنماذج التوليدية لاستكشاف سيناريوهات مستقبلية معقدة. ورغم ما يحمله هذا التطور من فرص واعدة، إلا أنه يضاعف في الوقت ذاته حجم المسؤولية الملقاة على عاتق الجهات المنظِّمة والمطوِّرة. ولذلك فإن الحوكمة الرشيدة لا يجب أن تكون عائقاً أمام الابتكار، بل شرطاً أساسياً لاستدامته وبناء الثقة المجتمعية فيه، وهو ما يستدعي التحوّل من ردود الفعل التنظيمية إلى التنظيم الاستباقي الداعم للابتكار في ذات الوقت، والذي يدمج الاعتبارات الأخلاقية والتنظيمية في جميع مراحل دورة حياة البيانات المصطنعة. فكلما زادت قدرة البيانات المصطنعة على تشكيل واقعنا، تعاظمت الحاجة إلى حوكمة أخلاقية تضع الإنسان في قلب الاهتمام، لضمان بقاء هذه التقنيات أدوات لتعزيز الرفاهية والعدالة والازدهار. إن ضمان تحقيق ذلك هو التحدي الأكبر والفرصة العظمى في آن واحد.

المراجع

1. European Parliament & Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). vol. L119 1–88 (2016).

2. Gartner. Gartner Data & Analytics Summit 2022 Orlando: Day 1 Highlights. https://www.gartner.com/en/newsroom/press-releases/2022-08-22-gartner-data-analytics-summit-2022-orlando-day-1-highlights (2022).

3. The Royal Society. *Synthetic Data - What, Why and How?* https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Synthetic_Data_Survey-24.pdf (2024).

4. Drechsler, J. & Haensch, A.-C. 30 Years of Synthetic Data. *Stat. Sci.* 39, 221–242 (2024).

5. Personal Data Protection Commission (PDPC), Singapore. *Proposed Guide on Synthetic Data Generation*. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/other-guides/proposed-guide-on-synthetic-data-generation.pdf (2024).

6. United Nations Economic Commission for Europe (UNECE). *Synthetic Data for Official Statistics: A Starter Guide*. https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf (2022).

7. Global Alliance for Genomics and Health (GA4GH). *When Are Synthetic Health Data Personal Data?* https://www.ga4gh.org/news_item/when-are-synthetic-health-data-personal-data/ (2024).

8. OECD.AI Policy Observatory. *Synthetic Data in Financial Services*. https://oecd.ai/en/wonk/synthetic-data-financial-services (2023).

9. European Data Protection Supervisor (EDPS). *Synthetic Data*. https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en (2023).

10. IBM. What Is Synthetic Data? https://www.ibm.com/think/topics/synthetic-data (2024).

11. Financial Conduct Authority (FCA), Synthetic Data Expert Group (SDEG). *Generating and Using Synthetic Data for Models in Financial Services: Governance Considerations*. https://www.fca.org.uk/publication/corporate/synthetic-data-models-financial-services-governance.pdf (2025).

12. Mendes, J. M., Barbar, A. & Refaie, M. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Front. Digit. Health* **7**, (2025).

13. Mantellassi, F. *Governance Implications of Synthetic Data in the Context of International Security*. https://unidir.org/wp-content/uploads/2024/12/UNIDIR_Governance_Implications_Synthetic_Data.pdf (2024).

14. Future of Privacy Forum (FPF), "Synthetic Content Report, (2024). https://fpf.org/wpcontent/uploads/2024/10/Synthetic_Content_Report_October_2024.pdf

15. Thakur, M. & Hausenloy, J. Opportunities and Challenges of Frontier Data Governance with Synthetic Data. in *ICLR 2025 Workshop on Human-AI Coevolution (HAIC)* (2025).

16. United Nations University (UNU). *Recommendations on the Use of Synthetic Data to Train AI Models*. https://collections.unu.edu/eserv/UNU:9480/Use-of-Synthetic-Data-to-Train-AI-Models.pdf (2024).

17. European Parliament & Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024).

18. European Parliament & Council of the European Union. Regulation (EU) 2022/868 of the European Parliament and of the Council on European data governance (Data Governance Act). vol. L152 1–44 (2022).

19. United Kingdom. Data Protection Act 2018. (2018).

20. Information Commissioner's Office (ICO). UK GDPR guidance and resources. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources (2025).

21. Office for National Statistics (ONS). Synthetic Data Policy. https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/syntheticdatapolicy (2023).
22. NHS England. Artificial Data. https://digital.nhs.uk/services/artificial-data (2024).
23. 117th U.S. Congress. H.R.847 - Advancing Synthetic data Ability and Practice Act (ASAP Act). https://www.congress.gov/bill/117th-congress/house-bill/847 (2021).
24. Utah State Legislature. S.B. 149: Artificial Intelligence Amendments. https://le.utah.gov/~2024/bills/static/SB0149.html (2024).
25. California State Legislature. Assembly Bill No. 2013 - Generative AI: Training Data Transparency. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013 (2024).
26. National Institute of Standards and Technology (NIST). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://www.nist.gov/itl/ai-risk-management-framework (2023) doi:10.6028/NIST.AI.100-1.
27. *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf (2024) doi:10.6028/NIST.AI.100-4.
28. Garfinkel, S. L., McCallister, E., Scholl, M. & Pillitteri, V. *De-Identifying Government Datasets*. https://doi.org/10.6028/NIST.SP.800-188 (2024) doi:10.6028/NIST.SP.800-188.
29. Near, J. P., Darais, D., Lefkovitz, N. & Howarth, G. S. Guidelines for Evaluating Differential Privacy Guarantees. National Institute of Standards and Technology (NIST) Special Publication (SP) 800-226 (2025). https://doi.org/10.6028/NIST.SP.800-226
30. U.S. Census Bureau. Survey of Income and Program Participation (SIPP) Synthetic Beta Data Product. https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html (2024).
31. Cyberspace Administration of China. Internet Information Service Deep Synthesis Management Provisions (Draft for Comment). (2022).
32. Cyberspace Administration of China (CAC). *Interim Measures for the Management of Generative Artificial Intelligence Services*. (2023).
33. Cheng, L. Exploring Synthetic Data – privacy, utility and fidelity. https://www.priv.gc.ca/en/blog/20221012/ (2022).
34. Office of the Privacy Commissioner of Canada (OPC). *A Pan-Canadian Descriptive Study of Privacy Risks from Synthetic Data-Generation Practices within the Evolving Canadian Legislative Landscape*. https://www.priv.gc.ca/en/opc-actions-and-decisions/research/funding-for-privacy-research-and-knowledge-translation/completed-contributions-program-projects/2021-2022/p_2021-22_10/ (2024).
35. IBM. *Unlocking AI Opportunities with the Responsible Generation and Use of Synthetic Data*. https://www.ibm.com/downloads/documents/us-en/1379d3325a5ae60f (2025).
36. OECD. *Emerging Privacy-Enhancing Technologies: Current Regulatory and Policy Approaches*. https://doi.org/10.1787/bf121be4-en (2023) doi:10.1787/bf121be4-en.
37. OECD. *Enhancing Access to and Sharing of Data in the Age of Artificial Intelligence*. https://doi.org/10.1787/23a70dca-en (2025) doi:10.1787/23a70dca-en.
38. IEEE Standards Association. Synthetic Data. https://standards.ieee.org/industry-connections/activities/synthetic-data/ (2024).
39. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). Information security, cybersecurity and privacy protection — Information security controls. https://www.iso.org/standard/86899.html (2022).

تـاريخ الوثيقـة | يوليو 2026م
نـــوع الوثيقـة | دراسات وتقارير
تصنيف الوثيقة | عام
رقـــم الوثيقـة | SDAIA-P149
رقـــم الإصدار | 1

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

AITTA
مركـــز دراســـات الذكـــاء الاصـــطناعـــي
Artificial Intelligence Think Tank

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