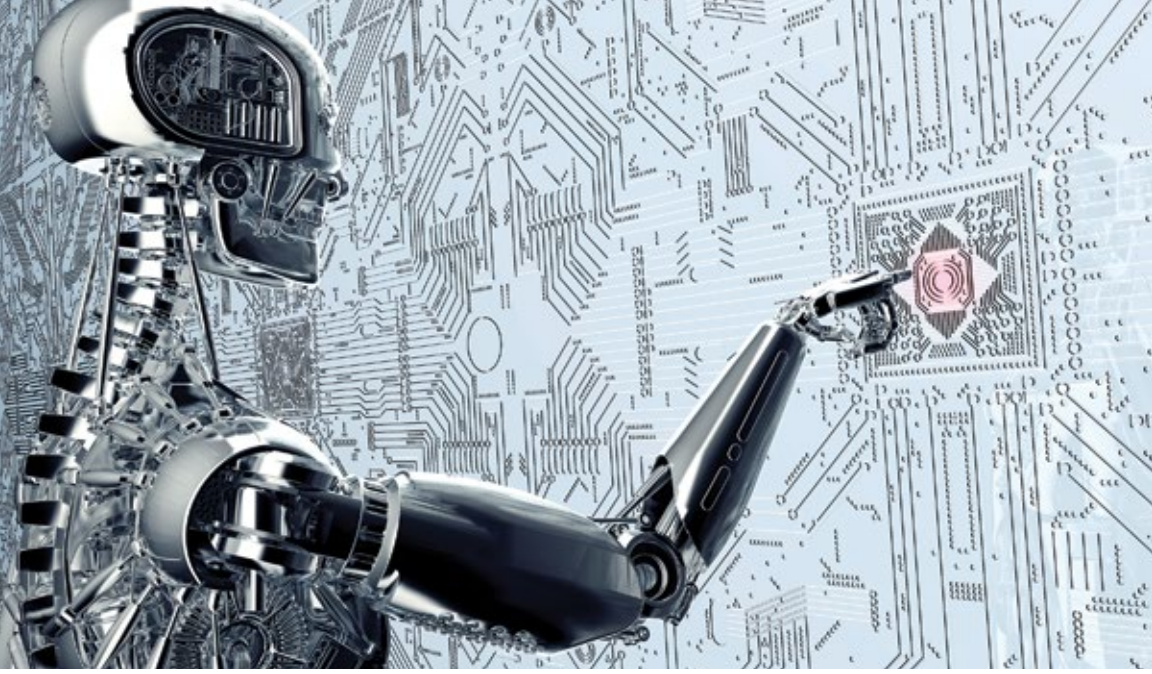


ذكاء اصطناعي بملامح بشرية

مخاطر التحيز والأخطاء في الذكاء الاصطناعي

RAND CORPORATION

ذكاء اصطناعي بملامح بشرية

مخاطر التحيز والأخطاء في الذكاء الاصطناعي

أوشونديه أوشوبا (Osonde Osoba) ووليام ويلسر الرابع (William Welser IV)

للحصول على مزيدٍ من المعلومات حول هذا المنشور، الرجاء زيارة الموقع الإلكتروني

www.rand.org/t/RR1744

بيانات فهرس النشر في مكتبة الكونغرس متاحة لهذا المنشور.
الرقم الدوليّ المعياريّ للكتاب: 978-0-8330-9763-7

نشرته مؤسسة RAND، سانتا مونيكا، كاليفورنيا

الغلاف: *the-lightwriter/iStock/Getty Images Plus*

مؤسسة RAND مؤسسة بحثية تقوم بتطوير الحلول للتحديات التي تواجه السياسات العامة للمساعدة في جعل المجتمعات في جميع أنحاء العالم أكثر أمنًا وأمانًا وصحة وازدهارًا. مؤسسة RAND هي مؤسسة غير ربحية، حيادية، وملتزمة بالصالح العام.

لا تعكس منشورات مؤسسة RAND بالضرورة آراء عملاء ورعاة الأبحاث الذين يتعاملون معها.

ادعم مؤسسة RAND
تبرع بمساهمة خيرية تُختصم من الضرائب على الصفحة
www.rand.org/giving/contribute

www.rand.org

تمهيد

تؤثر الخوارزميات ووكلاء الذكاء الاصطناعي (أو "وكلاء الذكاء الاصطناعي" مجتمعة) على كثير من جوانب الحياة مثل: قراءة المقالات الإخبارية، والحصول على الائتمان، واستثمار رأس المال، من بين جملة أمور أخرى. وبسبب كفاءتها وسرعتها، تتخذ الخوارزميات القرارات وتنفذ الإجراءات نيابة عن البشر في هذه المجالات وكثير غيرها. وعلى الرغم من هذه المكاسب، هناك مخاوف بشأن المكننة السريعة للوظائف (حتى الوظائف المعرفية، مثل الصحافة والطب الإشعاعي). مع ذلك، لا يُظهر هذا الاتجاه أي علامات على التراجع. ومع استمرار ازدياد الاعتماد على وكلاء الذكاء الاصطناعي، ما العواقب والمخاطر المترتبة على هذا الاعتماد؟ ومن الضروري بلوغ فهم أفضل للمواقف تجاه الخوارزميات وأشكال التفاعل معها، ولا سيما بسبب هالة الموضوعية والتنزه عن الخطأ التي تضفيها ثقافة اليوم على الخوارزميات. ويوضح هذا التقرير بعض أوجه القصور في اتخاذ القرار الخوارزمي، ويحدد المواضيع الرئيسية التي تدور في فلك مشكلة الأخطاء والتحيزات الخوارزمية (مثل تغذية البيانات والتأثير المتباين الخوارزمي)، ويدرس بعض المقاربات لمكافحة هذه المشاكل. هذا التقرير على قدر كبير من الأهمية لصناع القرار والمنفذين الذين يسعون إلى اكتساب فهم أفضل لكيفية تأثير نشر الذكاء الاصطناعي على أصحاب الشأن لديهم. وهذا من شأنه التأثير على مجالات مثل العدالة الجنائية، والأشغال العامة، وإدارة الرعاية الاجتماعية.

مشاريع مؤسسة RAND

مؤسسة RAND مؤسسة بحثية تقوم بتطوير الحلول للتحديات التي تواجه السياسة العامة للمساعدة في جعل المجتمعات في جميع أنحاء العالم أكثر أمنًا وأمانًا وصحة وازدهارًا. مؤسسة RAND هي مؤسسة غير ربحية، حيادية، وملتزمة بالصالح العام. تم توفير

التمويل لهذا المشروع من خلال المساهمات السخية للمجلس الاستشاري لمركز RAND للمخاطر والأمن العالميّ (CGRS). وقد تم إجراء البحث داخل المركز، كجزء من البرامج الدولية بمؤسسة RAND.

تعتبر مشاريع مؤسسة RAND وسيلة للاستثمار في حلول السياسات. وتدعم المساهمات الخيرية قدرتنا على تبنّي وجهة نظر بعيدة المدى، ومعالجة المواضيع الصعبة والمثيرة للجدل في كثير من الأحيان، وتبادل النتائج التي توصلنا إليها بطرق مبتكرة ومقنعة. وتستند نتائج أبحاث مؤسسة RAND وتوصياتها إلى البيانات والأدلة، وبالتالي فهي لا تعكس بالضرورة تفضيلات السياسات أو الاهتمامات الخاصة بعملائها أو جهاتها المانحة أو مؤيديها.

تم توفير التمويل لهذا المشروع من خلال المساهمات السخية للمجلس الاستشاري لمركز RAND للمخاطر والأمن العالميّ. وقد تم إجراء البحث داخل المركز، كجزء من البرامج الدولية بمؤسسة RAND.

كما يتم تقديم الدعم لهذا المشروع جزئيًا من الرسوم التي يتم تحصيلها من الأبحاث الممولة من الذين يتعاملون مع RAND.

مركز RAND للمخاطر والأمن العالميّ

يعتمد مركز RAND للمخاطر والأمن العالميّ على الخبرة الواسعة والمتعددة التخصصات لمؤسسة RAND لتقييم أثر الاتجاهات السياسية والاجتماعية والاقتصادية والتكنولوجية على المخاطر والأمن العالمي.

للحصول على مزيد من المعلومات حول مركز RAND للمخاطر والأمن العالميّ، قم بزيارة الموقع www.rand.org/international/cgrs أو الاتصال بمدير المركز (معلومات الاتصال متوفرة على صفحة المركز على شبكة الإنترنت).

المحتويات

الشكل

شكر وعرفان

نود أن نشكر أندرو باراسيليتي (Andrew Parasiliti) على رعاية هذه الدراسة من خلال مركز RAND للمخاطر والأمن العالميّ. كما نود أن نشكر الباحثين الزملاء الذين قدموا المشورة الصائبة أثناء عملنا على هذه الدراسة، ونخص بالشكر بارت كوسكو (Bart Kosko) وأنجيلا أوماهوني (Angela O'Mahony) وسارة نوفاك (Sarah Nowak) وجون ديفيس (John Davis). وأخيرًا نود أن نشكر المراجعين لدينا جون سيلي براون (John Seely Brown) وتيموثي آر. غولدن (Timothy R. Gulden) على آرائهم المتبصرة.

مقدمة

تؤثر الخوارزميات ووكلاء الذكاء الاصطناعي (أو "وكلاء الذكاء الاصطناعي" مجتمعة) على كثير من جوانب الحياة مثل: قراءة المقالات الإخبارية، والحصول على الائتمان، واستثمار رأس المال، بين جملة أمور أخرى. وبسبب كفاءتها وسرعتها، تتخذ الخوارزميات القرارات وتنفذ الإجراءات نيابة عن البشر في هذه المجالات وكثير غيرها. وعلى الرغم من هذه المكاسب، هناك مخاوف بشأن المكننة السريعة للوظائف (حتى الوظائف المعرفية، مثل الصحافة والطب الإشعاعي). مع ذلك، لا يُظهر هذا الاتجاه أي علامات على التراجع.[1]

ومع استمرار ازدياد الاعتماد على وكلاء الذكاء الاصطناعي، كذلك تزداد المخاطر! ومن الضروري بلوغ فهم أفضل للمواقف تجاه الخوارزميات وأشكال التفاعل معها، ولا سيما بسبب هالة الموضوعية والعصمة من الخطأ التي تضفيها ثقافة اليوم على الخوارزميات (Bogost, 2015). ماذا يحدث عندما نسمح لما أسماه غودارد ورودزاري ووايت (2012) "تحيز المكننة" بأن ينتشر ويستشري؟ قد تكون الأخطاء الخوارزمية في الأنظمة الترفيهية التجارية محدودة التأثير (مثل إرسال أحدهم إلى موعد غرامي مع شخص غير متوافق معه). ولكن الخوارزميات الخاطئة في مجالات البنية التحتية (شبكات الكهرباء) أو أنظمة الدفاع أو الأسواق المالية يمكن أن تشكِّل مخاطر شديدة على الأمن العالمي. يوضح "الانهيار المفاجئ" لعام 2010 مدى الضعف الذي يمكن أن يوصلنا إليه اعتمادنا على وكلاء الذكاء الاصطناعي (Nuti et al., 2011).

1 إن الاتساق الإجرائي حُجة مؤيدة لنموذج اتخاذ القرار المدعوم. فاستخدام الخوارزميات يحد من تأثير عملية اتخاذ القرار الذاتية أو الاعتباطية. لكن سيترون (2007) (Citron، ص 1252) ذهبت إلى أن الاستخدام واسع النطاق لأدوات المساعدة في المكننة والقرار الخوارزمي جعل من الأنظمة الرقمية "صانعات القرار الرئيسية في السياسات العامة" بدلًا من أن تكون أدوات مساعدة للقرار في بعض نواحي القانون الإداري. كما طرحت سيترون أسئلة تتعلق بمراعاة أصول الإجراءات القضائية الواجبة: ربما لا توفر القرارات المتخذة عن طريق الخوارزميات سوى مجالات محدودة للاستئناف أو استدراك الإنصاف المشروع.

وقد أشار مكتب البيت الأبيض لسياسة العلوم والتكنولوجيا إلى أن اعتمادنا المتزايد على الأدوات الاصطناعية المبهمة يشكل تهديدًا للخصوصية والحقوق المدنية والاستقلالية الفردية، محذرًا من "إمكانية ترميز (إدخال) التمييز في القرارات الممكننة" (المكتب التنفيذي للرئيس، 2016، ص 45) ومناقشًا سمات المشكلة ودراسات الحالة في مجالات مثل إعداد التقارير الائتمانية وفرص العمل والتعليم والعدالة الجنائية (انظر أيضًا المكتب التنفيذي للرئيس، 2014). وبالتالي، من المهم تقييم مدى هذا التهديد وشدته.

وهدفنا هنا هو شرح المخاطر المرتبطة بالاعتماد الضعيف على الخوارزميات، ولا سيما عندما تعمل بشكل ضمني أو صريح كوسيط للحصول على الخدمات والفرص (مثل الخدمات المالية، والائتمان، والإسكان، والتوظيف). إن القرارات الخوارزمية ليست منصفة بشكل تلقائي بحكم أنها منتجات لعمليات معقدة، كما أن الاتساق الإجرائي للخوارزميات لا يكافئ الموضوعية. ويصف ديديو (2015) (DeDeo، ص 1) هذه المسألة بإيجاز: «قد تكون [الخوارزميات] مثالية من ناحية الرياضيات، ولكن تثير الإشكاليات من الناحية الأخلاقية». وعلى الرغم من أن صنع القرار البشري أيضًا حافل بالتحيزات المماثلة التي قد تبديها وكلاء الذكاء الاصطناعي، تصبح مسألة المساءلة أكثر غموضًا عند الاعتماد على العوامل الاصطناعية.

أما باقي التقرير فسيتخذ الهيكل التالي: يعرّف الفصل الثاني مفهوم الخوارزمية ويتناوله بالشرح، ثم ننتقل إلى الخوارزميات المعقدة التي تنحى منحىً غير صحيح أو غير منصف. وسينصب تركيزنا الأساسي على تأثير الأدوات الاصطناعية في مجالات المجتمع والسياسة. وينصرف الفصل الثالث بعيدًا عن أمثلة معينة ليحلل القضايا الكامنة وراء مشكلة خوارزميات سيئة الأداء. وسنقترح مجموعة مختارة من سُبل العلاج لاستعادة قدر من المساءلة عن عمليات اتخاذ القرارات الخوارزمية. وهذا يشمل العمل الذي تم مؤخرًا على التعلم الآلي الذي يتّسم بالعدل والمساءلة والشفافية. وفي الفصل الرابع، سنختتم التقرير ببعض الملاحظات والتوصيات حول كيفية تحسين فهم ومعالجة تحديات التحيز الخوارزمي.

الخوارزميات: تعريف وتقويم

كثيرًا ما لا يتضح للناس طبيعة الخوارزميات التي تتحكم بأجزاء كبيرة من حياتهم. وحاليًا يزداد اعتماد صناع القرار ومحللي السياسات على الخوارزميات وهم يحاولون اتخاذ قرارات فعالة في الوقت المناسب في عالم غني بالبيانات، ويتضمن استخدامهم للخوارزميات كوسائل مساعدة على اتخاذ القرارات (أو وكلاء الذكاء الاصطناعي بشكل أعم) تفاصيل هامة ولكنها ليست ذات صلة بالقرار. وهذه فائدة قوية لوسائل المساعدة الخوارزمية على اتخاذ القرارات. فالخوارزميات التي تعمل بشكل صحيح تطلق العنان لقدرة صناع القرار المعرفية من أجل المداولات الهامة الأخرى.[1]

ولكن غموض الخوارزميات يجعل من الصعب الحكم على صحة الأداء وتقويم المخاطر وتقييم العدالة في التطبيقات الاجتماعية. كما يمكن للغموض أن يحجب الفهم السببي وراء القرارات. قد تكون هذه المسائل غير ضارة إذا كانت الخوارزميات (شبه) مُنزهة عن الخطأ، ولكن معظم الخوارزميات لا تضمن الدقة سوى ضمانات احتمالية وحسب، وهذا في أفضل السيناريوهات الممكنة حيث يتم تطبيق النماذج والخوارزميات الصحيحة بشكل مناسب، مع توافر أفضل النوايا تجاه "إتمام" البيانات. ونادرًا ما يكون لدى مصممي الخوارزميات ومستخدميها رفاهية هذه السيناريوهات المثالية، إذ عليهم أن يعتمدوا على الافتراضات التي يمكن أن تفشل وتؤدي إلى نتائج غير متوقعة.[2]

1 إن الاتساق الإجرائي حُجة مؤيدة لنموذج اتخاذ القرار المدعوم. فاستخدام الخوارزميات يحد من تأثير عملية اتخاذ القرار الذاتية أو الاعتباطية. لكن سيترون (2007) (Citron، ص 1252) ذهبت إلى أن الاستخدام واسع النطاق لأدوات المساعدة في المكننة والقرار الخوارزمي جعل من الأنظمة الرقمية "صانعات القرار الرئيسية في السياسات العامة" بدلًا من أن تكون أدوات مساعدة للقرار في بعض نواحي القانون الإداري. كما طرحت سيترون أسئلة تتعلق بمراعاة أصول الإجراءات القضائية الواجبة: ربما لا توفر القرارات المتخذة عن طريق الخوارزميات سوى مجالات محدودة للاستئناف أو استدراك الإنصاف المشروع.

2 على سبيل المثال، يقول سلمون (2012) (Salmon) أن الانهيار المالي في عام 2008 كان نتيجة لفرط الاعتماد على نموذج غير دقيق لارتباط مخاطر التقصير في سداد الديون، وهو "رابطة جاوس".

يسهل إثبات قابلية حدوث الخطأ في الخوارزميات، ويشمل ذلك الأخطاء الخوارزمية في الأنظمة وليس فقط أخطاء عدم الدقة الإحصائية الملازمة للعديد من الخوارزميات. ويوجد العديد من الأمثلة في التطبيقات الموجهة للسياسات العامة. وتعتبر أداة "غوغل لتوجهات الانفلونزا" مثالًا ملموسًا على الأخطاء الضخمة حيث تشتهر بتكرار الخطأ في تشخيص اتجاهات الانفلونزا على المستوى الوطني (Lazer et al., 2014). تم إعداد الكثير من خوارزميات تقييم المخاطر بناءً على نماذج احتمالية خاطئة، وقد فشلت في الاستجابة المسبقة بشكل صحيح قبل الانهيار المالي في الولايات المتحدة عام 2008 (Salmon, 2012). في إحدى المدن تم تطبيق خوارزميات كان الهدف منها أن تكشف بشكل مثالي الحفر والأخاديد الموجودة في الشوارع بناءً على بيانات يتم جمعها من مستخدمي الهواتف الذكية. لكن التوزيع الديموغرافي لمستخدمي الهواتف الذكية في ذاك الوقت أسفر عن عدم رصد بعض الحفر والأخاديد في الشوارع ما تسبب في نقص الخدمات لبعض المجتمعات المحرومة (Crawford, 2013). وكان هذا من شأنه أن يؤدي إلى حرمان المواطنين الأقل ثراء من الحصول على خدمات الصيانة في تلك المدينة. بينما قررت مدينة أخرى أن تستخدم المقاربات الخوارزمية لتوجيه أنشطة إنفاذ القانون لديها. وكان التبرير لذلك أن الخوارزميات البوليسية التنبؤية تكون أكثر موضوعية لأنها لا تعتمد سوى على "معادلات موضوعية متعددة المتغيرات" وليس على القرارات البشرية الشخصية (مقتبس في Tett, 2014). وفي معرض تقريره على تطبيق آخر في مجال العدالة الجنائية، أظهر أنغوين وآخرون (Angwin et al. (2016) التحيز المنهجي في إحدى خوارزميات تقييم المخاطر الجنائية المستخدمة في جلسات النطق بالحكم في مختلف جميع أنحاء الولايات المتحدة.

تعريف الخوارزميات

من المفيد أن نبحث بشكل دقيق في ماهية الخوارزميات ونحن نمضي قدماً. لقد شهد هذا المفهوم بعض التغيير على مر القرون. فقد كان أبو عبد الله بن موسى الخوارزمي – العالم المسلم من القرون الوسطى الذي اشتُق اسم الخوارزمية من اسمه – مهتماً بابتكار إجراءات تفصيلية للتوصل إلى الحلول الحسابية للمعادلات (Arndt, 1983). وقد طرح ألونزو تشيرتش وآلان تورنغ (Alonzo Church and Alan Turing) (Turing, 1937a; Turing, 1937b) مفهومي "قابلية الحوسبة" و"الدوال الحسابية" لصياغة فكرة الخوارزمية. وانتهى التعريف إلى كونها متتالية محدودة من الأوامر الدقيقة القابلة للتنفيذ في الأنظمة الحاسوبية (والتي تشمل على سبيل المثال لا الحصر، العقل البشري).

وربما يذكّرنا هذا بإجراءات الحفظ عن ظهر قلب المدعومة بالعقل، مثل وصفات تحضير أطباق الطعام أو الخطوات المتبعة لحساب عبء ضريبتك الفيدرالية. ويقودنا تعريف تشيرتش وتورنغ مباشرة إلى المفهوم الشائع للخوارزميات بوصفها كود برمجي لمعالجة الأرقام ببراعة.

كان الراحل مارفن مينسكي (Marvin Minsky) (1961) وغيره من رواد الذكاء الاصطناعي (من أمثال، جون مكارثي (John McCarthy) وفرانك روزنبلات (Frank Rosenblatt)) الذين سارت أعمالهم على نهج أعمال تشيرتش وتورنغ يفكرون في جانب آخر من الخوارزميات: وهو تمكين أنظمة الحوسبة بقدرة الذكاء! ومن أبرز الملامح الرئيسية المميزة للذكاء القدرة على التكيف أو التعلم الاستقرائي من "التجربة" (أي البيانات). وقد أسفرت جهودهم عن تشكيل خوارزميات "التعلم" لتدريب أنظمة الحوسبة على التعلم و/أو إنشاء نماذج داخلية مفيدة عن العالم. كما تتألف هذه الخوارزميات من إجراءات حوسبية تسلسلية محفوظة على المستوى متناهي الصغر. والفرق هنا هو أن الخوارزميات تعالج الأرقام ببراعة عبر نماذج رياضية ثابتة بل وتحدّث أداءها بشكل متكرر بناءً على النماذج المصممة استجابةً لتجاربها (البيانات المُدخلة) ومقاييس الأداء.[3] ومع ذلك تستمر الصعوبة البالغة لمشكلة التعلم.[4] فقد حاولت الكثير من الخوارزميات الأولية أن تحاكي السلوكيات البيولوجية.[5] وكان الهدف الأكبر (ولا يزال) هو ابتكار ذكاء اصطناعي مستقل قادر على استخدام مثل خوارزميات التعلم المتطورة هذه لينافس الذكاء البشري المرن أو يتفوق عليه. وغالبًا ما يُطلق على مثل هذه الأنظمة في النقاشات المعاصرة أنظمة "الذكاء الاصطناعي العام". وتُظهر النجاحات التجارية – مثل الانتصار الأخير لبرنامج "ألفا غو AlphaGo" الذي أنتجته غوغل (Silver et al., 2016) وروبوت الدردشة بتقنيات الذكاء الاصطناعي المتقدم "تاي

3 يقول فاليانت (Valiant) (2013) إنّ التطوّر بحد ذاته هو نوع من خوارزميات التعلم التي تعمل بصورة متكررة على تكييف الخصائص الاجتماعية والبيولوجية لتحسين مقياس أداء صلاحية التوالد والتناسل.

4 مثُعرف شكلة تعلم التمييز بين الحقيقة والكذب بالاعتماد على الخبرة باسم "مشكلة الاستقراء". والسؤال المحوري هنا هو ما مدى تسويغ تطبيق التعميمات المستندة إلى خبرات سابقة محدودة على سيناريوهات جديدة؟ أولى الفلاسفة هذه المشكلة الكثير من التفكير. وأعرب ديفيد هيوم (David Hume) عن قلقه بشكل خاص إزاء استخدام الاستقراء للتعلم عن السببية (Hume, 2000, Sec. VII). ويفسر برتراند راسيل (Bertrand Rus-sell) هذه النقطة بمثال الدجاجة التي تعلمت التعرف على المُزارع كونه سبب حصولها على وجباتها اليومية وذلك بالاعتماد استقرائيًا على خبرات سابقة طويلة، ولا يُتاح للدجاجة أي سبب منطقي لتتوقع أن المزارع سيكون سبب موتها في النهاية (Russell, 2001).

5 كان هناك موجة تفاعل أولية بين رواد الذكاء الاصطناعي وعلماء النفس (السلوكيين والفيسيولوجيين على حد سواء) في محاولة لفهم كيفية تعلم الحيوانات لسلوكيات جديدة.

Tay" الذي صممته مايكروسوفت (Lee, 2016) – مدى التقدم الذي حققه هذا المسار من الأبحاث.

ويشكل الجزء الأكبر من عمل رواد الذكاء الاصطناعي أساس خوارزميات التعلم الآلي التي تقوم عليها معظم الأنظمة المُمكننة المستخدمة اليوم. وتركز هذه الأنظمة المُمكننة عادة على التعلم على إيجاد حلول لمهامٍ "بسيطة" مثل الكلام التلقائي والتعرف على الصور. والمصطلح الشائع الذي يُطلق على هذه الأنظمة هو "الذكاء الاصطناعي المحدود". ويُعزى نجاحها جزئيًا إلى الزيادة المهولة في القدرة الحاسوبية المتاحة لتنفيذ خوارزمياتها وتوسيعها. فعلى سبيل المثال، يُشكّل عملها أساسًا لأحدث أساليب التعلم العميق المستخدمة في التعرف الحديث على الصور والكلام.[6]

وتُشكّل ثورة "البيانات الضخمة" المستمرة حافزًا قويًا يشجّع على استخدام خوارزميات التعلم على نطاقٍ واسعٍ. وتوفر البيانات الضخمة (Brown, Chui, and Manyika, 2011) التدفق المستمر للبيانات متعددة الأنماط اللازمة لاستخلاص الرؤى القيمة عبر خوارزميات التعلم. ولن ينمو هذا التدفق سوى عندما تصبح الكائنات أكثر تشابكًا (على سبيل المثال، في "إنترنت الأشياء") لإنتاج المزيد من البيانات. والسبيل المستدام الوحيد لفهم الحجم المطلق للبيانات التي يتم إنتاجها يوميًا وتنوعها هو تطبيق خوارزميات قوية.

يميل تصورنا الثقافي للخوارزميات إلى تضخيم كل ما يتعلق بالخوارزميات من إجراءات الحوسبة المعماة (أي الحوسبيات الساكنة) وحتى التعلم المتقدم المُمكنن وإجراءات الاستدلال المستخدمة في حاسوب واطسون (Watson) الذي ابتكرته شركة آي بي إم (Ziewitz, 2016)، وقد ذهب أيان بوجوست (2015) (Ian Bogost) إلى أن هذا التصور الثقافي للخوارزميات ما هو إلا اجتزاء مغلوط يشجّع الأشخاص غير المتخصصين على التعامل مع الخوارزميات على أنها كيانات غامضة غير موضحة المعالم تكاد تتصل بعالم اللاهوت. كما يتم اعتبار الكثير من الخوارزميات الرئيسية المؤثرة في الحياة العامة على أنها ملكية خاصة أو من الأسرار التجارية، وهذه الأستار المسدلة من السرية لا تشجّع الخطاب العام المستنير.

إن هذا الفهم الغامض غير المستنير للخوارزميات يعيق الخطاب العام الذكي حول أوجه القصور فيها. على سبيل المثال، كيف نناقش مسائل حول صحة الخوارزميات في ضوء التنوع الواسع لها؟ فمن جهة، تعتبر صلاحية خوارزمية الحوسبة المعماة مرتبطة بمدى صحة تنفيذها، فعلى سبيل المثال، هل تقوم خوارزمية اقتراح النصائح الحسابية

6 يشير مفهوم "التعليم العميق" إلى استخدام العديد من طبقات المعالجة الخفية في معماريات التعلم الآلي الاتصالية (المكونة من عناصر معالجة متصلة) (مثل الشبكات العصبية).

بتنفيذ عمليات ضرب وجمع النسب المئوية بشكل صحيح؟ وهل تقوم خوارزمية حساب الأعباء الضريبية بإيلاء الاعتبار الواجب للدخل الخاضع للضريبة مع تطبيق القواعد المناسبة بما يتماشى مع قانون الضرائب؟ وهل قامت خوارزمية الفرز بعملية الفرز الدقيقة لمجموعة البيانات بأكملها أم أهملت أجزاءً منها؟ هذه الأسئلة تتناول المفاهيم الملموسة التي يمكن أحيانًا التحقق منها بشكل موضوعي.

إلا أن صلاحية الخوارزمية التعليمية ذات طبيعة مختلفة بعض الشيء، فهي تتعلق بصحة تنفيذها للمهام (وهذا ما يركز عليه مصممو الخوارزميات) وبصحة سلوكها المتعلَم المكتسَب (وهذا ما يهم المستخدمين العاديين). فلنأخذ كمثال معاصر روبوت الدردشة بتقنية الذكاء الاصطناعي "تايTay " الذي أطلقته شركة مايكروسوفت؛ فقد تم تطبيق الخوارزميات الذي يرتكز عليها "تاي" وتمكينها بالشكل الصحيح لتتحادث بطريقة بشرية مقنعة مع مستخدمي تويتر. ولم تظهر الاختبارات المكثفة في البيئات الخاضعة للمراقبة أية علامات تحذيرية. وقد كانت أبرز المزايا السلوكية له قدرته على التعلم والاستجابة لميول وتفضيلات المستخدمين من خلال استيعاب بياناتهم. وقد مكنت هذه الميزة مستخدمي تويتر من التلاعب بسلوك "تاي" مما تسبّبَ بإصدار الروبوت لسلسلة من العبارات البذيئة (Lee, 2016). فإن تجربته وبياناته لم تأخذا بعين الاعتبار الحداثة في سياقها الجديد.

لا يقتصر هذا النوع من الثغرات على هذا المثال فحسب، فخوارزميات التعلم عادةً ما تكون غير حصينة أمام خصائص بياناتها التدريبية. وهذه سمة من سمات هذه الخوارزميات: القدرة على التكيُّف استجابةً للمدخلات المتغيرة، لكن التكيُّف الخوارزمي في سياق الاستجابة للبيانات المُدخلة يفتح أيضًا بابًا لهجوم المستخدمين من أصحاب الأغراض الخبيثة. وتعتبر ثغرات تغذية البيانات في خوارزميات التعلم موضوعًا متكررًا.

الخوارزميات "سيئة الأداء": استعراض موجز

بما أن وكلاء الذكاء الاصطناعي يؤدون دورًا أكبر في عمليات اتخاذ القرار، ينبغي إيلاء مزيد من الاهتمام إلى تأثيرات وكلاء الذكاء الاصطناعي القابلة للخطأ أو سوء الأداء.

فالأدوات الاصطناعية كما يبدو من اسمها ليست بشرية؛ وعادة ما يتطلب الحكم الأخلاقي توفر عنصر الاختيار أو التعاطف أو الفاعلية لدى القائم بالفعل. لا يمكن أن توجد أي منظومة أخلاقية ذات معنى لدى الأدوات الاصطناعية؛ فسلوكها تحدده

سببيًا المواصفات البشرية.[7] إن مصطلح "الخوارزميات سيئة الأداء" هو مجرد استعارة للإشارة إلى وكلاء الذكاء الاصطناعي التي تؤدي نتائجها إلى تبعات غير صحيحة أو غير منصفة أو خطيرة.

إن تاريخ هذه الأدوات الاصطناعية سيئة الأداء يمتد على الأقل إلى وقت ظهور أنظمة الحوسبة السائدة. وقد ناقشت باتيا فريدمان (Batya Friedman) والفيلسوفة هيلين نيسنباوم (Helen Nissenbaum) (1996) مخاوف التحيز في استخدام أنظمة الحاسوب لأداء مهام متنوعة مثل الجدولة، ومطابقة الوظائف مع المرشحين، وتوجيه الرحلات الجوية، والمساعدة القانونية المُمكننة للهجرة. للوهلة الأولى بدا أن نقاش فريدمان ونيسنباوم يدور حول استخدام أنظمة الكمبيوتر، ولكن مقالتهم النقدية كانت تستهدف الإجراءات التي اعتادت هذه الأنظمة استخدامها من أجل توليد نتائجها، وهي الخوارزميات. وأفادت تحليلات فريدمان ونيسنباوم بوجود سلوك غير منصف أو متحيز في هذه الخوارزميات واقترحت إطارًا منهجيًا للتفكير في مثل هذه التحيزات.

كتب فريدمان ونيسنباوم (1996) عن نظام حجز الرحلات الجوية "بيئة الحجوزات شبه المُمكننة للأعمال" (SABRE) والذي قامت شركة الخطوط الجوية الأمريكية برعايته (انظر أيضًا Sandvig et al., 2014). وقد وفر نظام "بيئة الحجوزات شبه المُمكننة للأعمال" خدمة غيرت ملامح القطاع، فقد كان النظام أحد أوائل الأنظمة الخوارزمية التي توفر قوائم الرحلات ومعلومات المسارات لرحلات خطوط الطيران في الولايات المتحدة. إلا أن سلوك الفرز الافتراضي للمعلومات استفاد من سلوك المستخدم النموذجي لإنشاء تحيز مناهض للمنافسة في النظام يصب في مصلحة الجهة الراعية له.[8] فدائمًا ما كان نظام "بيئة الحجوزات شبه المُمكننة للأعمال" يعرض للوكلاء رحلات من الخطوط الجوية الأمريكية في الصفحة الأولى، حتى عندما كان لدى شركات الطيران الأخرى رحلات أرخص أو أقل في ساعات التوقف لنفس الاستعلام. وكثيرًا ما كان يتم إبعاد الرحلات الجوية غير المفضلة إلى الصفحات التالية والتي نادرًا ما كان يصل إليها الوكلاء. واضطرت شركة الخطوط الجوية الأمريكية إلى جعل نظام

7 تدور العديد من السجالات بشأن المسؤولية في الأنظمة المُمكننة حول هذا السؤال: ما درجة استقلالية الذكاء الاصطناعي الكافية للحد من المسؤولية الأخلاقية لمشرفي النظام من البشر عن عواقب أعمال الذكاء الاصطناعي؟ على سبيل المثال، إلى أي مدى تكون شركة مثل غوغل أو فيسبوك أو تسلا مسؤولة عن التأثيرات غير المتوقعة من الدرجة الثانية أو الثالثة أو أعلى لاستخدام أنظمتها المُمكننة؟ كيف نعيّن حدود إمكانية التنبؤ بنظام هو بالأساس (في الوقت الراهن على الأقل) مبهم؟ ويوضح مثال التشهير الذي نعرضه عن غوغل لاحقًا أن المحاكم القضائية بدأت تتعامل مع مثل هذه الأسئلة حول حدود المسؤولية القانونية.

8 أطلق موظفو الخطوط الجوية الأمريكية على هذه الممارسة اسم عِلم الشاشة.

"بيئة الحجوزات شبه الآلية للأعمال" أكثر شفافية بعد تسليط إجراءات مكافحة الاحتكار الضوء على هذه المسائل.

كما قام فريدمان ونيسنباوم (1996) بدراسة تاريخ الخوارزمية من أجل "البرنامج الوطني لمطابقة المقيم الطبي" الذي يطابق المقيمين الطبيين مع المستشفيات في جميع أنحاء الولايات المتحدة. ويبدو أن قواعد تخصيص الخوارزمية التي تبدو منصفة في أدائها قد تحيزت لتفضيلات المستشفيات على حساب تفضيلات المقيمين ولتفضيلات المقيمين العازبين على حساب المقيمين المتزوجين.[9] كما قام فريدمان ونيسنباوم (1996) بدراسة تاريخ الخوارزمية من أجل "البرنامج الوطني لمطابقة المقيم الطبي" الذي يطابق المقيمين الطبيين مع المستشفيات في جميع أنحاء الولايات المتحدة. ويبدو أن قواعد تخصيص الخوارزمية التي تبدو منصفة في أدائها قد تحيزت لتفضيلات المستشفيات على حساب تفضيلات المقيمين ولتفضيلات المقيمين العازبين على حساب المقيمين المتزوجين.9 كما بحث فريدمان ونيسنباوم في "برنامج قانون الجنسية البريطانية" الذي تم تصميمه لترميز قانون المواطنة البريطاني. ويشتمل القانون نفسه على مشاكل تمس الإنصاف، وقد توارثت أي خوارزمية ملتزمة بتنفيذ القانون هذه المشاكل وبالتالي قامت بتضخيمها. وكانت النقطة الأكثر إثارة للاهتمام هي أن برنامج قانون الجنسية البريطاني قدم ردودًا رسمية أخفت خيارات قانونية معنية موجودة بالقانون عن المستخدمين غير الخبراء. وكانت إجابات البرنامج صحيحة من الناحية الإجرائية، ولكن ترجمة القانون إلى خوارزمية دقيقة أفقده خصائص دقيقة هامة. إن الأنظمة التي تناولها فريدمان ونيسنباوم هي الأنظمة الكبيرة الشاملة التي كانت شائعة في الأيام الأولى للحوسبة الشخصية والإنترنت. وقد وسّع النمو الضخم للإنترنت وقاعدة مستخدمي الحواسيب الشخصية من نطاق هذه المشاكل. فقد بدأت الخوارزميات تتوسط المزيد من تفاعلاتنا مع المعلومات. ويُعتبر غوغل المثال النموذجي في هذا الصدد، فقد كانت خوارزميات البحث ووضع الإعلانات لدى غوغل تستهلك كميات هائلة من البيانات التي ينتجها المستخدمون كي تتعلم تحسين الخدمة للمستخدمين (سواء المستخدمين العاديين أو الجهات الإعلانية). وكانت هذه الأنظمة بعض أوائل الأنظمة التي كشفت عن نتائج خوارزميات التعلم أمام الاستهلاك الشخصي واسع الانتشار.

9 يشير هذا المثال إلى خوارزمية المطابقة المستخدمة في "البرنامج الوطني لمطابقة المقيم الطبي" قبل أن يقوم ألفين روث (Alvin Roth) بتغييرها في منتصف التسعينيات (Roth, 1996). وكان هذا الإجراء للمطابقة تجسيدًا لخوارزمية المطابقة المستقرة التي وضعها لأول مرة ديفيد غيل (David Gale) ولويد شابلي (Lloyd Shapley) (Gale and Shapley, 1962). وهي مستقرة بمعنى أنها لا تحابي المستشفى ولا المقيم على حساب الآخر نظرًا لترتيب الأفضليات المعلن للمجموعات بأكملها. ولكن خلافًا للمزاعم الأولية، أدى البرنامج إلى مطابقات مستقرة تضمن للمستشفيات أفضل الخيارات المقبولة ولكنها لا تضمن للطلاب سوى خيارات مقبولة وحسب (وفي بعض الأحيان أقل الخيارات قبولًا).

كانت لاتانيا سويني (Latanya Sweeney) ونيك دياكوبولوس (Nick Diakopoulos) من الرواد في دراسة سوء الأداء في أنظمة غوغل (Swee-ney, 2013; Diakopoulos, 2013; Diakopoulos, 2016). وقد كشف عملهم عن أمثلة على التشهير الخوارزمي في عمليات بحث غوغل وإعلاناته. وساق دياكوبولوس مثالًا نموذجيًا لهذا التشهير الخوارزمي حيث تتعلم برامج الإكمال التلقائي لمحرك البحث – والتي يغذيها تلقيم مستمر من الاستعلامات التاريخية للمستخدمين – كيفية تكوين ارتباطات تشهيرية أو متعصبة غير صحيحة عن أشخاص أو مجموعات من الأشخاص.[10] وأظهرت سويني أن هذه الارتباطات السلبية المتعلَّمة تؤثر على إعلانات غوغل المستهدفة. وفي المثال الذي ساقته، أسفر مجرد البحث عن أنواع معينة من الأسماء عن الإعلان عن خدمات العدالة الجنائية، مثل سندات الكفالة أو التحقق من السجلات الجنائية. وشملت أمثلة دياكوبولوس ارتباطات تشهيرية على الدوام لعمليات البحث المتعلقة بقضايا المتحولين جنسيًا.

تعتبر الدراسات على غرار دراسات سويني ودياكوبولوس أعمال أصيلة في مجال البيانات والصحافة الخوارزمية المتنامي. وتؤرِّخ المزيد من المقالات الإخبارية والبحثية أخطاء الخوارزميات التي تؤثر على مجالات مختلفة من حياتنا، أثناء الاتصال بالإنترنت وعدم الاتصال به. فاز حاسوب الذكاء الاصطناعي "واطسون Watson" الذي طورته شركة آي بي إم (IBM) باختبار جيوباردي (Jeopardy)، ومن المعروف أنه كان يجب تصحيح عادة إظهاره كلمات نابية بعد استيعاب خوارزميات التعلم به لبعض البيانات غير الأخلاقية (Madrigal, 2013). كما وردت تقارير عن تأثيرات خوارزميات توجيه حركة المرور على أنماط حركة المرور في المناطق الحضرية في تطبيق Waze (Bradley, 2015). ويصف أحد الكتب الكاشفة شذوذ البيانات والخوارزميات الكامنة وراء خدمة موقع المواعدة الغرامية الشهير "أوك كيوبيد OkCupid" (Rudder, 2014). وفي الآونة الأخيرة، كشف متعاقدو فيسبوك السابقون عن أن خوارزمية اتجاهات التلقيمات الإخبارية الجديدة في فيسبوك كانت في الواقع نتيجة لمدخلات ذاتية من فريق بشري (Tufekci, 2016).

وبدأ باحثون آخرون يكتبون عن آثار الخوارزميات في الحوكمة والسياسات العامة والقضايا الاجتماعية الشائكة. ويجب على وكلاء الذكاء الاصطناعي أن تتعامل مع التحديات التي تفرضها طبقة أخرى من التعقيدات والمخاطر في هذه الميادين. فيمكن أن يكون لإساءة السلوك هنا عواقب بعيدة المدى بين الأجيال تطال جميع شرائح المجتمع.

10 تحمِّل ألمانيا الآن غوغل المسؤولية الجزئية عن صحة اقتراحات الإكمال التلقائي (Diakopoulos, 2013).

عرض سيترون (2007) (Citron) تقريرًا عن كيف أن وصول صنع القرار الخوارزمي إلى المجالات القانونية يحرم المواطنين من حق الحصول على الإجراءات القضائية الواجبة. ومؤخرًا، أفاد أنجوين وآخرون (2016) (Angwin et al.) بوجود تحيز منهجي شديد في خوارزمية واسعة الاستخدام لتقييم المخاطر الجنائية في جلسات إصدار الأحكام في جميع أنحاء البلاد.

كتب سيترون (Citron) وباسكوال (2014) (Pasquale) عما أسمياه "المجتمع المتدرج" وهفواته. وما يقصدانه بمصطلح "المجتمع المتدرج" هو الحالة الراهنة التي تقوم فيها خوارزميات مبهمة غير منتظمة وأحيانًا خفية بمنح درجات رسمية للسمعة الفردية والتي تتدخل في الحصول على الفرص. وتشمل هذه الدرجات: الائتمان، وصحيفة الحالة الجنائية، وقابلية التوظيف. وقد ركز سيترون وباسكوال بشكل خاص على كيفية انتهاك هذه الأنظمة للتوقعات المعقولة للإجراءات القضائية الواجبة، ولا سيما توقعات الإنصاف والدقة وتوافر سبل الانتصاف. كما ذهبا إلى أن إحراز درجات الائتمان في الخوارزمية لم يقلل من التحيز والممارسات التمييزية. بل على العكس من ذلك، تضفي هذه الدرجات شرعية على التحيز الموجود فعليًا في البيانات (والبرمجيات) المستخدمة لتدريب الخوارزميات. واتبع باسكوال (2015) منهج بحثي مماثل مع تقرير شامل حول الاستخدام المكثف للخوارزميات غير المنتظمة في ثلاثة مجالات محددة، وهي: إدارة السمعة (مثل إحراز درجات الائتمان)، وسلوك محركات البحث، والأسواق المالية.

وكتب باروكاس (Barocas) وسيلبست (2016) (Selbst) مقالًا مؤثرًا حديثًا يتناول السؤال الجوهري حول إن كان يمكن للبيانات الضخمة أن تثمر عن سلوك عادل أو محايد في الخوارزميات.[11] وأجاب الباحثان على هذا السؤال بالنفي القاطع دون طرح إصلاح لكيفية تطبيق البيانات الضخمة والخوارزميات المرتبطة بها. يستعير باروكاس وسيلبست (وغيرهما من الباحثين في هذا المجال) مصطلح "التأثير المتباين" من الفقه القانوني الذي تم طرحه أول مرة في ستينيات وسبعينيات القرن الماضي لاختبار عدالة ممارسات التوظيف.[12] ويستخدم المؤلفان المصطلح للدلالة على المساوئ المنهجية التي تفرضها وكلاء الذكاء الاصطناعي على المجموعات الفرعية بناءً على الأنماط المتعلَمة

11 يستخدم باروكاس وسيلبست البيانات الضخمة كمصطلح شامل للإشارة إلى مجموعات البيانات الهائلة والأساليب الخوارزمية لتحليل هذه البيانات.

12 في قضية غريغز ضد شركة ديوك باور (1971)، قضت المحكمة العليا الأمريكية ضد استخدام شركة ديوك باور أنواع معينة من اختبارات ومتطلبات التوظيف التي لم يكن لها علاقة بالوظيفة. وقد تكون هذه الاختبارات غير مؤذية في ظاهرها، ولكن عند إجراء فحص دقيق عن كثب، نجد هذه الاختبارات "مثيرة للكراهية" وتحرض على التمييز على أساس العرق.

من خلال الإجراءات التي تبدو معقولة وغير تمييزية في ظاهرها. وقد استخدم غاندي (Gandy) (2010) مصطلح "التمييز العقلاني" للإشارة إلى مفهوم متطابق. وكان يناقش ضرورة فرض قيود تنظيمية على أنظمة دعم القرار لمعالجة الآثار الخارجية السلبية الجامحة (مثل المساوئ التراكمية) التي تعززها هذه الأنظمة.

ناقش باروكاس ونيسنباوم (2014) كيف أن الخوارزميات والبيانات الضخمة تتحايل أيضًا على أي ضمانات قانونية تتعلق بالخصوصية قد نشأنا نتوقعها. وإن الضمان القياسي ضد نتائج التأثير المتباين للخوارزمية هو إخفاء حقول البيانات الحساسة (مثل الجنس والعرق) عن خوارزميات التعلم. ولكن الدراسات السابقة حول تقنيات إعادة التعريف الحديثة تقر بأن خوارزميات التعلم يمكنها أن تعيد بناء الحقول الحساسة بشكل ضمني وتستخدم هذه المتغيرات البديلة المستنبطة بالاحتمالات من أجل التصنيف التمييزي (DeDeo, 2015; Feldman et al., 2015). ولا تزداد قوة أساليب الاستنباط سوى عندما تتم إضافة المزيد من مجموعات البيانات إلى قاعدة التدريب (Ohm, 2010). وهذا يطرح مشكلة على مستوى التنظيم؛ يمكن سن تشريعات ضد الاستخدام الصريح للمعلومات المحمية (مثل العرق والجنس في قانوني تكافؤ فرص العمل والإسكان العادل)،[13] ولكن من الصعب إصدار تشريع ضد استخدام المعلومات المستنبطة بالاحتمال. وأفاد باسكوال (2015) بأن وكلاء دمج البيانات تستغل بالفعل هذه الثغرة التنظيمية.

بدأ مؤخرًا مصممو وباحثو الخوارزميات العمل على المقاربات التقنية للتحقق من التأثيرات المتباينة للخوارزميات و/أو التخلص منها. وقدم فيلدمان وآخرون (Feldman et al.) (2015) مقاربة للتحقق من أن إحدى خوارزميات التصنيف عادلة وفقًا للمعايير القانونية الأمريكية. ويدخل إجراء التصحيح لديهم تعديلات للحفاظ على الرتبة في البيانات المُدخلة للسيطرة على التأثير المتباين. وقد عرض ديديو (DeDeo) (2015) طريقة لتعديل مُخرجات إحدى الخوارزميات لفك ارتباط مخرجاتها عن المتغيرات المحمية. وقد استخدم دورك وآخرون (Dwork et al.) (2012) بعض رؤى دورك حول الخصوصية (Dwork, 2008a; Dwork, 2008b) لابتكار إطار نظري لخوارزميات التصنيف العادل. وتبحث هذه المقاربة في مقاييس التشابه العادل المراعي للسياق من أجل المقارنة بين الأفراد وتصنيفهم بغض النظر عن شمولهم في الفئة المحمية.

13 قانون الإسكان العادل الوارد في الباب الثامن من قانون الحقوق المدنية لعام 1968 والذي تنص عليه المجموعة 42 من القوانين الأمريكية 3606-3504.

دراسة حالة: وكلاء الذكاء الاصطناعي في نظام العدالة الجنائية

يلجأ نظام العدالة الجنائية في الولايات المتحدة بصورة متزايدة إلى أدوات خوارزمية. وتساعد وكلاء الذكاء الاصطناعي على تخفيف عبء إدارة مثل هذا النظام الضخم. ولكن أي تحيز خوارزمي منهجي في هذه الأدوات سيزيد من مخاطر وقوع الأخطاء والمساوئ التراكمية.

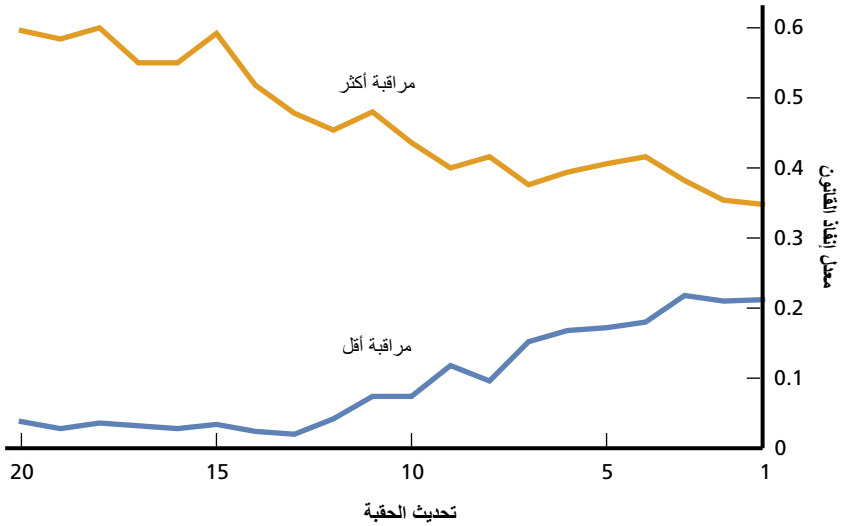
نتناول أولًا استخدام الخوارزميات في مرحلة المحاكمة وإطلاق السراح المشروط. وقد قدم أنجوين وآخرون (2016) تقريرًا حول نظام تقييم المخاطر الجنائية المتعلقة بتنميط إدارة الجناة الإصلاحية للعقوبات البديلة (COMPAS) والذي طوّرته شركة نورثبوينت. ويُستخدم هذا البرنامج في جلسات الحكم وإطلاق السراح المشروط في جميع أنحاء البلاد. وقدم أنجوين وآخرون قصصًا تبين أن النظام يسيء تمثيل مخاطر معاودة الإجرام لدى مختلف المدانين بالقضايا. تشير هذه القصص أولًا إلى تحيز عنصري منهجي في تقدير المخاطر؛ فقد كان يجري تصنيف المدانين السود تصنيفًا أعلى من المدانين من غير السود، حتى عندما ارتكب المدانون غير السود جرائم أشد خطورة. ويُتبع المؤلفون هذا التلميح بتحليل لبيانات تنميط إدارة الجناة الإصلاحية للعقوبات البديلة ومعاودة الإجرام من مقاطعة بروارد بولاية فلوريدا.

وقد اكتشف لارسون وآخرون (2016) (Larson et al.)، في عرضهم تفاصيل التحليل الإحصائي في أنجوين وآخرون (2016)، أن:

> المتهمون السود كانوا أكثر عرضة لسوء التصنيف من المتهمين البيض بمقدار الضعف في بحث ارتفاع احتمالات معاودة الإجرام العنيف، بينما تمت إساءة تصنيف معاودي الإجرام من البيض كمصدر خطر متدني بنسبة 63.2% في كثير من الأحيان عن المتهمين السود.

كما تلجأ أجهزة الشرطة إلى أدوات خوارزمية من أجل أعمال الشرطة التنبؤية وتخصيص الموارد. ونقدم مثالًا على المحاكاة يوضِّح كيف أن خوارزمية مقبولة رياضيًا تؤدي إلى سلوك تنبؤي غير منصف لأعمال الشرطة. يبيّن الشكل (1) كيف أن خوارزمية صحيحة رياضيًا للعثور على المجرمين استنادًا إلى البيانات التاريخية للجرائم يمكن أن تؤدي إلى سلوك غير منصف. ويوضِّح هذا الشكل السلوك المحاكي لنظام مُمكنن لتوجيه جهود إنفاذ القانون في العثور على الحوادث الجنائية والاستجابة لها في جميع الشرائح السكانية.

الشكل 1
معدل أحداث إنفاذ القانون لكل حقبة: شريحتان فرعيتان من السكان، ونفس معدل الجريمة، واختلاف درجة الاحتراس



RAND *RR1744-1*

لنفترض أن لدينا مجموعة من السكان مقسمة بشكل طبيعي إلى فئات (مثل الموقع أو الجنس أو نوع الجريمة أو أي معيار آخر) وأن لدينا موارد محدودة لإنفاذ القانون لا يمكنها اكتشاف جميع الحوادث الإجرامية والاستجابة لها. بالتالي، بالنسبة لسكان يعانون من معدل جريمة معين، سنكتشف انتهاكات وسنطبق القانون باحتمالية معينة. ويحدد مقياس الاحتراس مقدار الاهتمام الذي نوليه في مختلف الفئات وبالتالي احتمال الاكتشاف الناجح للجرائم ومقاضاة مرتكبيها عندما تقع. وتتكيف درجة احتراس النظام باستخدام إجراء من إجراءات التعليم المنطقية من الناحية الرياضية. ويزيد النظام من درجة الاحتراس في المناطق التي تزداد فيها ملاحظة النزعة الإجرامية و/أو في المناطق التي يرتفع فيها نشاط إنفاذ القانون استنادًا إلى بيانات الإنفاذ المسجلة.[14]

إن الخوارزمية فعالة بشكل معقول ولكنها تؤدي بسهولة إلى نتائج غير منصفة. ويبين الرسم البياني في الشكل محاكاة لسلوكها الذي يسفر عن تواتر متزايد بشكل غير متكافئ لأحداث إنفاذ القانون في الشرائح السكانية الفرعية المختلفة. ومن شأن هذا

14 يتشابه قانون تعلم الخوارزمية مع الأفكار التي تقوم عليها مقاربة النوافذ المحطمة في إنفاذ القانون والتي طرحها ويلسون (Wilson) وكيلنج (1982) (Kelling).

السلوك أن يكون مقبولًا إذا كان لا يتحقق إلا عندما تكون المستويات الأساسية للسلوك الإجرامي غير متكافئة أيضًا. ولكن الشكل يُظهر أن هذا السلوك المتباين الخاص بشرائح السكان الفرعية لا يزال ممكن التحقق عندما تظل معدلات الجريمة الأساسية الخاصة بالشرائح السكانية الفرعية هي نفسها. ويترتب على ذلك أثر تزايد تجريم شرائح سكانية فرعية معينة والأخطر من ذلك توليد المزيد من البيانات "الموضوعية" لدعم قرارات إنفاذ القانون المتحيزة في المستقبل.

إن الخوارزمية الموصوفة هي أكثر من مجرد أداة توضيحية، فهي تحاكي تأثير المراقبة غير المنصفة والمجردة من المبادئ التي تقوم بها الدولة استنادًا إلى السجلات التاريخية. وتفيد المحاكاة بأن ازدياد المراقبة التي تقوم بها الدولة ليس أداة محايدة، خاصةً إذا لم تطبق بشكل موحد. وعلى مستوى الشرائح السكانية ككل، يمكن أن يؤدي ذلك إلى "التجريم غير المنصف" حيث يكون للمجرمين من مختلف الخصائص الديموغرافية احتمالات مختلفة بشكل منهجي للاعتقال وأحكام بالسجن متفاوتة الشدة. وتطرح حركة "لون المراقبة" الحديثة أن المراقبة التي تقوم بها الدولة في الولايات المتحدة لم تُطبق بشكل مُنصف. ويرى بعض فقهاء القانون أن التجريم غير المنصف بات هو الموقف السائد في الولايات المتحدة، وغالبًا ما يتم تبريره استنادًا إلى سجلات تاريخية للجريمة، كما كان الحال في نظام تنميط إدارة الجناة الإصلاحية للعقوبات البديلة. وقد عثرت وزارة العدل الأمريكية (2016) على مزيد من الأدلة على هذه المراقبة وهذا التجريم غير المنصفين في التحقيق الذي أجرته مؤخرًا في جهاز شرطة بالتيمور.

ويمكن تطبيق خوارزمية مماثلة على مشكلة إيجاد مواقع صالحة للتعدين أو التنقيب عن النفط. إن سلوك الخوارزمية "غير المنصف" – الذي سيعطي مزيدًا من الاهتمام للمناطق ذات السجل التاريخي لاكتشاف النفط – يمكن اعتباره سمة (سيتم تركيز الاهتمام وتوجيه الموارد بشكل مثمر) وليس خللًا في هذا السياق. ويكمن جزء من الفارق في أن حلول مسائل السياسات العامة غالبا ما تحتاج إلى مراعاة مقاييس الجودة الأخرى التي تستقي معلوماتها من المبادئ الاجتماعية (التي تكون أحيانًا غامضة أو غير محددة) مثل الإنصاف أو العدل. وفي هذه الحالة، نتوقع أن يكون إنفاذ القانون عادلًا بمعنى أن يتناسب نشاط إنفاذ القانون مع النشاط الإجرامي عبر مختلف الفئات.

التركيز على المشكلة: العوامل وتدابير المعالجة

توضح الأمثلة في الفصول السابقة عددًا من زوايا تناول مشكلة تحيز الخوارزميات. الزاوية الأولى والأبسط هي مشكلة تغذية بيانات الخوارزمية؛ فنظرًا لمحدودية التوجيه البشري، فإن وكيل الذكاء الاصطناعي يعكس جودة البيانات التي يتعلم منها. ويؤدي التعلم المُمكنن من البيانات المتحيزة بطبيعتها إلى نتائج متحيزة. وتحاول خوارزميات الوكيل استخراج أنماط من البيانات تتمتع بمُدخلات بشرية محدودة خلال عملية الاستخراج. إن التوجيه البشري المحدود يثبت موضوعية العملية المنفذة، ولكن توليد البيانات غالبًا ما يكون ظاهرة اجتماعية (مثل عمليات التفاعل على وسائل التواصل الاجتماعي والخطاب السياسي عبر الإنترنت) تتأثر بالتحيزات البشرية. ويعتبر تطبيق الخوارزميات الصحيحة إجرائيًا على البيانات المتحيزة وسيلةً مؤكدةً لتعليم وكلاء الذكاء الاصطناعي تقليد أي تحيز تحتويه البيانات. على سبيل المثال، تظهر الأبحاث التي أجريت مؤخرًا أن الأساليب المُمكننة المطبقة على اللغة تتعلم بالضرورة التحيزات البشرية المتأصلة في استخدامنا للغة (Caliskan-Islam, Bryson, and Narayanan, 2016).

وهذا يؤدي إلى تأثير متناقض إلى حد ما وهو أن الأدوات الاصطناعية – والتي تتعلم بشكل ذاتي من البيانات المستمدة من البشر – غالبًا ما ستتعلم التحيزات البشرية، سواء كانت جيدة أم سيئة. ويمكننا أن نطلق على ذلك "مفارقة وكلاء الذكاء الاصطناعي". ويوضح مثالا "واطسون" و "تاي" هذه النقطة جيدًا. كما تطرح سويني (2013) أمثلة متعددة على الأنظمة الإعلانية المستهدفة التي تسفر عن استنتاجات متحيزة وأحيانًا تشهيرية حول أشخاص بعينهم بسبب التحيزات المُتعلمة بشكل ذاتي من البيانات. ولهذه المفارقة تداعيات هامة على استخدام وكلاء الذكاء الاصطناعي في عصر البيانات الضخمة. إن تعقيد أنماط البيانات وضخامة حجم البيانات المتاحة يفرضان على وكلاء الذكاء الاصطناعي أن تتعلم بشكل أكثر ذاتية. وهذا يشير إلى أنه ينبغي للأفراد أن يتوقعوا مزيدًا من وكلاء الذكاء الاصطناعي تحاكي التحيزات البشرية.

وغالبًا ما تنطبق الزاوية الثانية لتناول مشكلة تحيز الخوارزميات عند التعامل مع المسائل السياسية أو الاجتماعية. وهذه هي صعوبة تعريف الأدلة التجريبية أو تحديد مبادئ توجيهية قوية. وغالبًا ما تكون الأدلة التجريبية أو معاييرنا للحكم على صحتها موجهة ثقافيًا أو اجتماعيًا، كما يوضح مثالا حاسوب "واطسون" من شركة آي بي إم وخاصية الإكمال التلقائي من غوغل. وسيكون على خوارزميات التعلم تحسين أدائها بما يحقق قبول المجتمع بالإضافة إلى أي مقاييس أداء تقوم هذه الخوارزميات بتحسينها داخليًا لأداء مهمة من المهمات. ويؤدي هذا التحسين المزدوج إلى الوقوع في معضلات. وفي الواقع، فإن العمل الحديث على الخوارزميات العادلة يُظهر أن هناك عادةً مفاضلة بين الدقة والعدل. إن فرض قيود العدل يمكن أن يعني إغفال أو تشويه المتغيرات المعلوماتية بشكل كبير. وهذا يمكن أن يُضعف قوة الاستنتاج الخوارزمي.

هناك زاوية أخرى لهذه لمشكلة وهي أن الأحكام في مجال السلوك الاجتماعي غالبًا ما تكون غامضة بدلًا من أن تكون معايير ثنائية واضحة المعالم.[1] وتتناول هذه الزاوية النقطة الثانية باستفاضة. وتُظهر الأمثلة التي عرضناها مسبقًا المعايير الثقافية الغامضة («لا تشتم أو تسب!»، «لا تشهد زورًا!»، «أعرض وجهة نظر متوازنة!») التي تؤثر على الحكم البشري على سلوك الخوارزمية الصحيح. ونحن قادرون على تعلم الانتقال بين العلاقات المعقدة والغامضة، مثل الحكومات والقوانين، وغالبًا ما نعتمد على التقييمات الذاتية للقيام بذلك. ويمكن للأنظمة التي تعتمد على الاستدلال الكمي (مثل معظم وكلاء الذكاء الاصطناعي) أن تحاكي التأثير ولكنها غالبًا ما تتطلب تصميمًا دقيقًا للقيام بذلك. ولعل تحقيق هذا يتطلب ما هو أكثر من مجرد الكمبيوتر وعلماء البيانات.

لقد تطور نظام آخر عبر القرون لحل القضايا المتعلقة بالسياسات والتي تخضع لمعايير اجتماعية غامضة وتقارير أو بيانات متضاربة: وهو القانون. وأشار غريملمان (Grimmelmann) ونارايانان (2016) (Narayanan) إلى أنه على الرغم من أن العملات المشفرة والعقود الخوارزمية ("الذكية") قد تتفوق في إنفاذ حقوق الملكية الثنائية، إلا أن حقوق الملكية في العالم الحقيقي غامضة ومثيرة للخلافات. وتنطبق مخاوف مماثلة على الخوارزميات؛ فما نعتبره سلوكًا خوارزميًا سليمًا لا يمكن تعريفه بدقة سوى في بعض الأحيان. وقد تطور القانون من أجل الفصل في هذه التعقيدات الغامضة.

---

[1] في هذا السياق لكلمة "غامض" معنى دقيق، حيث تشير إلى الخصائص والمجموعات التي تتمتع بحدود تعريف غير حاسمة. وهي تقوم على فكرة المنطق متعدد القيم (أي ليس الثنائي) والعضوية المحددة التي طرحها لأول مرة مفكرون مثل ماكس بلاك (Max Black) (في عمله على المجموعات الغامضة) ولطفي زاده (Lotfi Zadeh) (في عمله على المنطق الغامض). ولنسوق مثالًا ملموسًا، فكِّر في مجموعة درجات الحرارة التي تصف بها الطقس "بالدافئ". إن الحد الفاصل بين درجات الحرارة "الدافئة" و "غير الدافئة" غير محدد المعالم. وفي الأمثلة التي استعرضناها على منتجات الذكاء الاصطناعي التي تلفظت بألفاظ نابية، فإن الألفاظ النابية ليست محظورة تمامًا وليست مسموحة تمامًا؛ فثمة درجات متنوعة لقبول المجتمع بها.

كما يقر القانون الأمريكي بأن الإجراءات الصائبة ظاهريًا يمكن أن يكون لها تأثير عكسي ومتباين. وإن استيعاب مفهوم التأثير المتباين لا ينتشر إلا ببطء وسط مجتمع البحث الخوارزمي. وهناك مجموعة متزايدة من الأعمال المعنية بالأثر الاجتماعي والقانوني للبيانات والخوارزميات (Gangadharan, Eubanks, and Barocas, 2015). وتُظهر مجموعة متزايدة من الأدلة أن الخوارزميات لا تعامل الشرائح السكانية المتنوعة بشكل عادل ومنصف ذاتيًا بحكم كونها خوارزميات منطقية (Barocas and Selbst, 2016; DeDeo, 2015; Dwork et al., 2012; Feldman et al., 2015; Hardt, 2014).

العوامل التقنية الأخرى

إلى جانب العوامل التي سبق أن ناقشناها، تعزز العوامل التقنية الأخرى من تحيز الخوارزميات. وخوارزميات التعلم المُمكننة تعاني مشاكل في معالجة التفاوتات في حجم العينات. وهذه نتيجة مباشرة لحقيقة أن خوارزميات التعلم الآلي هي في الأصل وسائل إحصائية وبالتالي فهي تخضع للقوانين الإحصائية الخاصة بحجم العينة. وقد تواجه خوارزميات التعلم صعوبة في تسجيل تأثيرات ثقافية معينة عندما يكون السكان مقسمين إلى عدة شرائح متعددة. ويرتبط ذلك بمشكلة الاستدلال الإحصائي على بيانات التدريب غير المستقرة إلى حد كبير (ولا سيما عندما لا تمثل النماذج الافتراضية التأثيرات غير المستقرة).

تفاوت حجم العينة

إن خوارزميات التعلم الآلي هي وسائل تقدير إحصائية. وغالبًا ما تتفاوت مقاييس خطأ التقدير بشكل عكسي مع أحكام عينة البيانات. وهذا يعني أن هذه الوسائل ستكون عادةً أكثر عرضة للخطأ في فصول التدريب منخفضة التمثيل عن غيرها. كما أن خوارزمية تقدير الائتمان ستكون أكثر عرضة للخطأ مع الشرائح السكانية الفرعية التي ينخفض تمثيلها تاريخيًا في أسواق الائتمان. وقد عرض داوكن (2015) (Dwoskin) تصويرًا ملموسًا لهذا التأثير. فنظام وسم الصور (Tagging System) المُمكنن من ياهو (Yahoo) أسفر عن اختيارات عنصرية لوسم الصور بسبب عدم التجانس الديموغرافي في بياناته التدريبية. وللاطلاع على نقاش وافٍ حول التفاوت في حجم العينة، انظر Hardt, 2014.

وظائف المكافئة المخترقة

إن وظائف المكافئة في التعلم الآلي ونظرية الذكاء الاصطناعي مستعارة من علم النفس السلوكي، كما يتبين من أعمال بي. إف. سكينر (B. F. Skinner). وهذه الوظائف هي الوسائل الرئيسية التي تتعلم من خلالها نظم التعلم الاصطناعي الحالية ماهية السلوك الصحيح. فخلال عملية التعلم التي يقوم بها وكيل الذكاء الاصطناعي، تحدد وظيفة المكافئة مدى مكافئتنا أو عقابنا للأفعال والقرارات الصائبة أو الخاطئة. ثم تقوم خوارزميات التعلم بتكييف معاملات وسلوك الوكيل من أجل تعظيم المكافئة الإجمالية. وبالتالي، غالبًا ما يُختزل تصميم سلوك الذكاء الاصطناعي إلى وظائف المكافئة التي لا توفر سوى الكم الكافي من الحوافز. إلا أنه يمكن التلاعب بهذه المقاربة السلوكية إلى التعلم. على سبيل المثال، لنفترض وجود روبوت للتنظيف مُصمم من أجل التخلص من كميات القمامة؛ يمكن أن يحصل الروبوت على مكافئات عن طريق إيقاف مستشعراته البصرية بدلًا من التنظيف. يطلق أموداي (Amodei et al.) وآخرون على هذه العملية "اختراق المكافئة". فوظيفة المكافئة المُحددة بشكل سيء يمكن أن تؤدي إلى تأثيرات أو سلوكيات جانبية غير مرغوبة في أنظمة الذكاء الاصطناعي. كما أن اختراق المكافئة شاغل حاضر في الوقت الذي يكيِّف البشر سلوكياتهم مع التقويم الخوارزمي. فالبشر يتعلمون التلاعب بالخوارزميات مع انكشاف القدر الكافي من المعلومات (مثل، معرفة الإشارات الرخيصة غير المتصلة التي تضعها أنظمة تصنيف الائتمان في اعتبارها عند إجراء التصنيفات).

الاختلافات الثقافية

تعمل خوارزميات التعلم الآلي عن طريق اختيار الميزات البارزة (المتغيرات) في البيانات التي ترتبط بمختلف السلوكيات (Hardt, 2014). والسلوكيات المحملة برموز ثقافية قد تؤدي إلى سلوك غير منصف. يطرح هارت (Hardt) مثالًا على كيف أن الاختلافات الثقافية في قواعد إطلاق الأسماء تؤدي إلى الإبلاغ عن الحسابات ذات الأسماء غير التقليدية على منصات التواصل الاجتماعي.[2]

[2] هذه ظاهرة ثقافية غالبًا ما يُطلق عليها "صراعات الأسماء Nymwars". وتذهب منصات مثل تويتر وغوغل بلس وبليزارد إنترتينمنت (منصة تطوير ألعاب) إلى أن إطلاق الأسماء الحقيقية على الحسابات يساعد في الحفاظ على المظهر اللائق على شبكة الإنترنت. وبالتالي، تقوم هذه المنصات بالإبلاغ عن و/أو حذف الحسابات التي تحمل أسماء وهمية. وتعتمد عملية التمييز بين الأسماء الحقيقية والأسماء الوهمية بشكل مكثف على ممارسات إطلاق الأسماء الغربية التقليدية (Hardt, 2014; Boyd, 2012).

المتغيرات المصاحبة المحيرة

غالبًا ما يلجأ مصممو الخوارزميات إلى حذف المتغيرات الحساسة من البيانات التدريبية في محاولة منهم للتخلص من التحيز في النظام الناتج (Barocas and Selbst, 2016). وكثيرًا ما يقول مصممو الأنظمة إن «النظام لا يمكن أن يتأثر بالتحيز لأنه لا يضع في حسبانه بعض المتغيرات الحساسة». يناقش باروكاس وسيلبست (2016) الأسباب وراء أن إخفاء المتغيرات الحساسة غالبًا لا يحل المشكلة! فوسائل التعلم الآلي يمكن أن تستنتج عن طريق الاحتمالات متغيرات خفية، مثل استخدام الرمز البريدي لاستنتاج (كبديل احتمالي) الدخل. ولهذه القدرة تداعيات قوية على خصوصية البيانات وسرية الهوية. ويتحدث الباحثون اليوم عن أن توقعات المستخدمين التقليدية بشأن خصوصية البيانات وسرية الهوية لم تعد متاحة بعد الآن (Dwork, 2008a; Nara-yanan and Shmatikov, 2010; Ohm, 2010). وهذا بسبب أن خوارزميات التعلم الآلي الحديثة قادرة على إعادة التعرف على البيانات بسهولة وإحكام.

تدابير المعالجة

إن معالجة قصور وكلاء الذكاء الاصطناعي أو تنظيمهم سيتطلب على الأرجح المزج بين المقاربات التقنية وغير التقنية. وثمة جهود حديثة تجري لتطوير أساليب تعلم آلي عادلة ومسؤولة وشفافة. ونحن نقترح تعزيز هذه الأساليب باستخدام المقاربات الأقل تقنية.

المقاربات الإحصائية والخوارزمية

يوجد مجال متنامي يركِّز على التعلم الآلي العادل والمساءل والشفاف، بالعمل على مقاربات تقنية من أجل ضمان عدالة الخوارزميات أو التحقق من تأثير المتباين في خوارزميات التعلم الآلي وتصحيحه. اقترح دورك وآخرون (2012) استخدام مقاييس المسافة أو التشابه المعدَّلة عند التعامل مع بيانات الأشخاص. والهدف من مقاييس التشابه هو تطبيق قيود عدالة صارمة عند مقارنة الأشخاص في مجموعات البيانات. كما اقترح ساندفيج وآخرون (2014) (Sandvig et al.) عددًا من خوارزميات إجراءات التدقيق التي تقارن بين مخرجات الخوارزميات والسلوك المنصف المتوقع. فيمكن أن تكون عمليات التدقيق المعتمدة على الخوارزميات أكثر جدوى ودقة عندما تكون رموز وإجراءات الخوارزميات مفتوحة المصدر. طرح ديديو (2015) مقاربة خوارزمية لضمان أن نماذج التعلم الآلي تفرض الاستقلالية الإحصائية بين النتائج والمتغيرات

المحمية. علاوة على ذلك، طرح فيلدمان وآخرون (Feldman et al.) اختبارًا للتحقق من إن كانت الخوارزمية تنتهك قواعد التأثير المتباين القانونية (بموجب قانون الولايات المتحدة) أم لا. وهذا يوفر مقياسًا مستبصر اجتماعيًا لتمام الخوارزمية المنشود. كما اقترحوا أيضًا نموذج إحصائي لتصحيح أوجه عدم الإنصاف في خوارزميات التصنيف، إلا أنه لم يخلو من النقائص: فهذه المخططات ستتنازل غالبًا عن بعض من قدرتها التنبؤية في مقابل بلوغ العدالة.

خوارزميات الاستدلال السببي

يقوم كل من جوديا بيرل (Judea Pearl) (2009)، وليون بوتو وآخرون (Bottou et al.) (2013)، وباحثون غيرهم (Athey, 2015) بدراسة وسائل من أجل تزويد خوارزميات التعلم الآلي بالاستدلال السببي أو المغاير للحقائق، وذلك بشكل أعم وعلى مدار نطاق زمني أطول. وهذا المسعى على قدر عظيم من الأهمية لأن أنظمة الاستدلال السببي المُمكننة يمكن أن تقدم روايات سببية واضحة من أجل الحكم على جودة عملية صنع القرار الخوارزمي؛ إذ تُعتبر التبريرات السببية الدقيقة للقرارات الخوارزمية هي أكثر سجلات التدقيق موثوقية للحكم على الخوارزميات.

توضح إحدى القضايا غير المسبوقة التي حكمت فيها المحكمة العليا الأمريكية بعقوبة الإعدام مدى أهمية الاستدلال السببي في اتخاذ القرارات (قضية ماكليسكي ضد كيمب، 1987). فقد درس الباحث القانوني ديفيد بالدوس (David Baldus) استخدام الأساليب التجريبية الكمية في اختبار الإسراف في قرارات إصدار أحكام الإعدام في كاليفورنيا (Baldus et al., 1980). ثم طبَّق بالدوس تحليله على ولاية جورجيا في دراسته عام 1983 (Baldus, Pulaski, and Woodworth, 1983). وقد استعانت الدراسة بتحليلات إحصائية مضبوطة بدقة للبيانات المستمدة من الملاحظات حول أحكام الإعدام لتوضيح التأثير غير المتكافئ لأحكام الإعدام لولاية جورجيا.[3] وقد شمل تحليل بالدوس الشامل حوالي 230 متغير.

شهدت القضية المنظورة حضور خبراء الإحصاء الذين ناقشوا النتائج التي توصلت إليها دراسة بالدوس. بل أن مداولات المحكمة تضمنت نقاشات مطولة حول مفاهيم إحصائية تفصيلية، مثل التعددية الخطية.

وفي النهاية توصلت المحكمة إلى أن الحكم بالإعدام صحيح لأن الدراسة لم تبدِ أي تحيز متعمد في قضية ماكليسكي. وكان هذا هو القرار النهائي رغم التوضيح المفصَّل

[3] اكتشف بالدوس وبولاسكي و وودورث (1983) بإيجاز أن العرق زاد من احتمالات (بموجب عامل تضعيفي يبلغ 4.3) الحصول على حكم بالإعدام بالمقارنة مع أحكام الإدانة الأخرى في ولاية جورجيا.

للتفاوت العنصري بنسبة 4 إلى 1 في نتائج إصدار الحكم. إلا أن تبرير المحكمة كان: برغم صحة دراسة بالدوس، لم تثبت الدراسة أن العرق كان عامل سببي في الحكم الصادر بحق ماكليسكي.

إن أردنا الاعتماد على الخوارزميات في عملية اتخاذ القرار الذاتية، فيجب تزويدها بأدوات لمراجعة العوامل السببية خلف القرارات المتخذة. فالخوارزميات التي يمكن مراجعتها للتعرف على العوامل السببية ستقدم تفسيرات أو مسوغات أوضح لنتائجها. ويُعتبر هذا على قدر خاص من الأهمية لتبرير النتائج غير المتكافئة إحصائيًا.

الوعي والشفافية الخوارزمية

من أجل مكافحة تحيز الخوارزميات، سيكون من المفيد تثقيف الجمهور وتسليحه بالوعي بأن الخوارزميات يمكن أن تؤدي إلى نتائج غير منصفة. وهذا يختلف عن مطالبة المستخدمين بفهم آليات العمل الداخلية لكل الخوارزميات؛ فلن يكون ذلك مجديًا. فسيكون من المفيد غرس جرعة صحية من الشكوك المدروسة لدى المستخدمين بما يكفي لتقليل حجم تأثير تحيز المكننة. وثمة آمال كبيرة معقودة على هذا الأمر. فالوقت الطويل الذي نقضيه في التفاعل مع الخوارزميات عبر الواجهات قد يجعل خطأ الخوارزميات أكثر وضوحًا.

على سبيل المثال، عادةً ما يتشكك مستخدمو مواقع المواعدة الغرامية الإلكترونية (نسبة متزايدة بوتيرة سريعة من السكان) في نتائج خوارزميات مطابقة المواعيد الغرامية. كما أن الأعمال الصحفية والأفلام الوثائقية التي تناولت الانهيار المالي بسبب الرهن العقاري المجحف عام 2008 قد ساعدت على تغذية الشك الثقافي الصحي المطلع في كفاءة الخوارزميات المعقدة. فلننظر إلى التقارير الحديثة عن استياء الجماهير من تطبيق "سكيتش فاكتور SketchFactor" (Marantz, 2015). استخدم التطبيق بيانات محشودة جماعيًا وخوارزميات التجميع من أجل حساب درجة "سوء" الأحياء السكنية. وكان هناك رد فعل سبلي شديد إزاء هذا التطبيق بناءً على الحساسية الثقافية واحتمالات أن يؤدي إلى اعتداءات تمييزية. كانت الجماهير قادرة على التعبير عن مخاوف من أن درجات تطبيق "سكيتش فاكتور" الخوارزمية ستؤدي إلى ترميز التحيزات الثقافية الموجودة مسبقًا عن الأحياء السكنية.

ربما سيكون الجمع بين الوعي والشفافية مفيدًا جدًا في هذا الصدد. فعادة ما تشير الشفافية هنا إلى التأكد من أن أي خوارزمية مستخدمة هي ميسورة الفهم. مجددًا، من المستبعد أن يتحقق ذلك دائمًا. فسيكون من المجدي والمفيد مزيد من الإفصاح عن القرارات والإجراءات التي يتوسط فيها وكلاء الذكاء الاصطناعي. وحاليًا ينتظر المستخدمون مثل

هذه الإفصاحات كما يتضح من رد الفعل الغاضب إزاء ممارسات فيسبوك الجديدة في حفظ ومعالجة البيانات. فبالإضافة إلى إرضاء رغبات المستخدمين، تسمح الشفافية من خلال الإفصاحات للمستخدمين بمزيد من الاطلاع واستهلاك المعلومات بمزيد من التشكك.

مقاربات الأفراد

إن الأبحاث التقنية التي تُجرى على خوارزميات الذكاء الاصطناعي والتعلم الآلي لا تزال في مهدها. كما أن مسائل التحيز والأخطاء النظامية في الخوارزميات تتطلب منهجًا مختلفًا من مصممي الخوارزميات وعلماء البيانات. فغالبًا ما يكون هؤلاء الممارسين مهندسين وعلماء لا يتمتعون بالقدر الكافي من الاطلاع على قضايا السياسات العامة والمجتمع. وكثيرًا ما تكون الخصائص الديموغرافية التي يدمجها مصممو الخوارزميات بعيدة عن التنوع. فمصممو الخوارزميات يتخذون العديد والعديد من خيارات التصميم، قد يكون لبعضها تداعيات بعيدة المدى. بالتالي، إن وعي مطوري الخوارزميات بالتنوع سيساعد في تحسين الحساسية من مشاكل التأثير المتباين الممكنة.

من زاوية أخرى، ينبغي وجود جرعة صحية من القيود التنظيمية من أجل تحقيق التوازن مع التوجه نحو تدابير معالجة التحيز في الخوارزميات. فأي نوع من العلاج سيتطلب أن تلتزم الخوارزميات بصورة أوثق بالقيمة الاجتماعية. فما هذه القيم؟ ومن سيقررها؟ فمع استغراق المجتمع أكثر وأكثر في هذا المجال، سينبغي تناول مسائل حرية التعبير، والرقابة، والإنصاف، وغيرها من المعايير الأخلاقية المقبولة.

الخاتمة

تناولنا في تقريرنا تحدي التأثير المتباين للخوارزميات، وأسباب توقعنا مشاهدة توسع في الاعتماد على الخوارزميات، وأفضل الخيارات للتخفيف من المخاطر المستقبلية. وستستمر مخاطر الخطأ والتحيز في الخوارزميات والذكاء الاصطناعي طالما يؤدي وكلاء الذكاء الاصطناعي أدوارًا بارزة أكثر من أي وقت مضى في حياتنا دون تنظيمهم.

عادةً ما تتخذ الاستجابة إلى وكلاء الذكاء الاصطناعي غير الخاضعين للتنظيم 3 أشكال عامة: تجنب الخوارزميات تمامًا، أو إضفاء الشفافية على الخوارزميات الأساسية، أو تدقيق مُخرجات الخوارزميات. من المستحيل تجنب الخوارزميات تمامًا؛ وثمة بضعة خيارات أخرى متاحة للتعامل مع طوفان البيانات الحالي. تتطلب شفافية الخوارزميات غرس مزيد من الوعي بين الجماهير القادرة على استيعاب الخوارزميات. لكن التطورات الحديثة في التعلم الاتصالي العميق تشير إلى أن الخوارزميات ربما لا تزال على درجة من التعقيد تحول دون الاستفادة من هذه الأفكار، حتى وإن كان بوسعنا تفكيك إجراءات الخوارزمية.

يذهب عمل كريستيان ساندفيج (Christian Sandvig) الأخير إلى أن الخيار الأخير – تدقيق الخوارزميات – ينبغي أن يكون الخيار الذي سنعتمد عليه في المستقبل (Sandvig et al., 2014). وبعض أنواع التدقيق تتجاهل آليات العمل الداخلية لوكلاء الذكاء الاصطناعي وتحكم عليها طبقًا لعدالة نتائجها. وهذا أشبه بحكمنا غالبًا على البشر: من خلال تبعات مُخرجاتهم (القرارات والأفعال)، وليس بناءً على محتوى أو براعة قواعدهم من الرموز (الأفكار). سيستفيد صنّاع السياسات أكبر فائدة من هذا الخيار، كما أنه سيرسي المعيار للحكم على أخلاقيات وكلاء الذكاء الاصطناعي من منظور العواقب والتبعات. وبالتالي سيكون التنظيم أيسر كثيرًا من هذا الإطار.

لعل نقاشات مثل نقاشنا هذا تتطرق أحيانًا إلى إضفاء صفات بشرية على وكلاء الذكاء الاصطناعي! هل الآلات بصدد التفكير مثلنا؟ وكيف يمكن أن نحكم عليها ونوجهها؟

إن التقدم الحالي في مجال وكلاء الذكاء الاصطناعي ربما سيجعل هذا الإضفاء للصفات البشرية عليها أقرب إلى الوضع الطبيعي السائد. ولعل هذا سيسفر عن منافع غير متوقعة مثل تعزيز فهم الجماهير لوكلاء الذكاء الاصطناعي بوصفها غير منزهة عن التحيزات، مثلها مثل البشر.

الاختصارات

AI	الذكاء الاصطناعي
COMPAS	تنميط إدارة الجناة الإصلاحية للعقوبات البديلة
IBM	شركة آلات الأعمال الدولية (آي بي إم)
SABRE	بيئة الحجوزات شبه المُمكننة للأعمال

المراجع

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete Problems in AI Safety," Ithaca, N.Y.: Cornell University Library, 2016. As of February 2, 2017: https://arxiv.org/abs/1606.06565

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks," *ProPublica*, May 23, 2016. As of December 5, 2016: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arndt, A. B., "Al-Khwarizmi," *Mathematics Teacher*, Vol. 76, No. 9, 1983, pp. 668–670.

Athey, Susan, "Machine Learning and Causal Inference for Policy Evaluation," *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, August 10–13, 2015, pp. 5–6.

Autor, David, "The Polarization of Job Opportunities in the U.S. Labor Market: Implications for Employment and Earnings," Washington, D.C.: Center for American Progress and the Hamilton Project, April 2010.

Baldus, David C., Charles Pulaski, and George Woodworth, "Comparative Review of Death Sentences: An Empirical Study of the Georgia Experience," *Journal of Criminal Law and Criminology,* Vol. 74, No. 3, Autumn 1983, pp. 661–753.

Baldus, David C., Charles A. Pulaski, George Woodworth, and Frederick D. Kyle, "Identifying Comparatively Excessive Sentences of Death: A Quantitative Approach," *Stanford Law Review*, Vol. 33, No. 1, November 1980, pp. 1–74.

Barocas, Solon, and Helen Nissenbaum, "Big Data's End Run Around Procedural Privacy Protections," *Communications of the ACM,* Vol. 57, No. 11, 2014, pp. 31–33.

Barocas, Solon, and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review,* Vol. 104, 2016, pp. 671–732.

Bogost, Ian, "The Cathedral of Computation," *Atlantic*, January 15, 2015. As of May 27, 2016:
http://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/

Bottou, Léon, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson, "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising," *Journal of Machine Learning Research,* Vol. 14, No. 1, 2013, pp. 3207–3260.

Boyd, Danah, "The Politics of Real Names," *Communications of the ACM*, Vol. 55, No. 8, 2012, pp. 29–31.

Bradley, Ryan, "Waze and the Traffic Panopticon," *New Yorker*, June 2, 2015. As of December 2, 2016:
http://www.newyorker.com/business/currency/waze-and-the-traffic-panopticon

Brown, Brad, Michael Chui, and James Manyika, "Are You Ready for the Era of 'Big Data'?" *McKinsey Quarterly,* Vol. 4, No. 1, October 2011, pp. 24–35.

Caliskan-Islam, Aylin, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Necessarily Contain Human Biases," Ithaca, N.Y.: Cornell University Library, 2016. As of February 2, 2017:
https://arxiv.org/abs/1608.07187

Citron, Danielle Keats, "Technological Due Process," *Washington University Law Review,* Vol. 85, No. 6, 2007, pp. 1249–1313.

Citron, Danielle Keats, and Frank A. Pasquale, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review,* Vol. 89, 2014.

Crawford, Kate, "Think Again: Big Data," *Foreign Policy,* Vol. 9, May 10, 2013.

DeDeo, Simon, "Wrong Side of the Tracks: Big Data and Protected Categories," Ithaca, N.Y.: Cornell University Library, May 28, 2015. As of March 7, 2017:
https://arxiv.org/pdf/1412.4643v2.pdf

Diakopoulos, Nicholas, "Algorithmic Defamation: The Case of the Shameless Autocomplete," Tow Center for Digital Journalism website, August 6, 2013. As of February 14, 2017:
http://towcenter.org/algorithmic-defamation-the-case-of-the-shameless-autocomplete/

———, "Accountability in Algorithmic Decision Making," *Communications of the ACM,* Vol. 59, No. 2, 2016, pp. 56–62.

Dwork, Cynthia, "An ad Omnia Approach to Defining and Achieving Private Data Analysis," in Francesco Bonchi, Eleena Ferrari, Bradley Malin, and Yücek Saygin, eds., *Privacy, Security, and Trust in KDD*, New York: Springer, 2008a, pp. 1–13.

———, "Differential Privacy: A Survey of Results," *International Conference on Theory and Applications of Models of Computation*, New York: Springer, 2008b, pp. 1–19.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, "Fairness Through Awareness," *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, Cambridge, Mass., January 8–10, 2012, pp. 214–226.

Dwoskin, Elizabeth, "Social Bias Creeps into New Web Technology," *Wall Street Journal*, August 21, 2015. As of December 5, 2016: http://www.wsj.com/articles/computers-are-showing-their-biases-and-tech-firms-are-concerned-1440102894

Equal Employment Opportunity Act of 1972. As of March 8, 2017: https://www.eeoc.gov/eeoc/history/35th/thelaw/eeo_1972.html

Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values*, Washington, D.C., May 2014. As of February 3, 2017: https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

———, *Preparing for the Future of Artificial Intelligence*, Washington, D.C., October 2016. As of February 3, 2017: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, "Certifying and Removing Disparate Impact," *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, August 10–13, 2015, pp. 259–268.

Friedman, Batya, and Helen Nissenbaum, "Bias in Computer Systems," *ACM Transactions on Information Systems,* Vol. 14, No. 3, July 1996, pp. 330–347.

Gale, David, and Lloyd S. Shapley, "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, Vol. 69, No. 1, January 1962, pp. 9–15.

Gandy, Oscar H., Jr., "Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems," *Ethics and Information Technology,* Vol. 12, No. 1, March 2010, pp. 29–42.

Gangadharan, Seeta Peña, V. Eubanks, and S. Barocas, *Data and Discrimination: Collected Essays*, Washington, D.C.: Open Technology Institute, 2015.

Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt, "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," *Journal of the American Medical Informatics Association,* Vol. 19, No. 1, January–February 2012, pp. 121–127.

Griggs v. *Duke Power Co.*, 401 U.S. 424, 1971.

Grimmelmann, James, and Arvind Narayanan, "The Blockchain Gang," *Slate*, February 16, 2016. As of December 5, 2016:
http://www.slate.com/articles/technology/future_tense/2016/02/bitcoin_s_blockchain_technology_won_t_change_everything.html

Hardt, Moritz, "How Big Data Is Unfair: Understanding Sources of Unfairness in Data Driven Decision Making," *Medium*, September 26, 2014. As of December 5, 2016:
https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Hume, David, *An Enquiry Concerning Human Understanding: A Critical Edition*, Tom L. Beauchamp, ed., Vol. 3, New York: Oxford University Press, 2000.

Jaimovich, Nir, and Henry E. Siu, "The Trend Is the Cycle: Job Polarization and Jobless Recoveries," paper, Cambridge, Mass.: National Bureau of Economic Research, 2012.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, May 23, 2016. As of December 6, 2016:
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani, "Google Flu Trends Still Appears Sick: An Evaluation of the 2013–2014 Flu Season," paper, Rochester, N.Y.: Social Science Electronic Publishing, Inc., March 13, 2014. As of February 2, 2017:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2408560

Lee, Peter, "Learning from Tay's introduction," blog, Microsoft website, March 25, 2016. As of December 5, 2016:
http://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.0001v8vtz3qddejwq702cv2annzcz

Madrigal, Alexis C., "IBM's Watson Memorized the Entire 'Urban Dictionary,' Then His Overlords Had to Delete It," *Atlantic*, January 10, 2013. As of December 5, 2016:
http://www.theatlantic.com/technology/archive/2013/01/ibms-watson-memorized-the-entire-urban-dictionary-then-his-overlords-had-to-delete-it/267047/

Marantz, Andrew, "When an App Is Called Racist," *New Yorker*, July 29, 2015. As of December 5, 2016:
http://www.newyorker.com/business/currency/what-to-do-when-your-app-is-racist

McCleskey v. *Kemp*, 481 U.S. 279, 1987.

Minsky, Marvin, "Steps Toward Artificial Intelligence," *Proceedings of the IRE,* Vol. 49, No. 1, 1961, pp. 8–30.

Narayanan, Arvind, and Vitaly Shmatikov, "Myths and Fallacies of Personally Identifiable Information," *Communications of the ACM,* Vol. 53, No. 6, June 2010, pp. 24–26.

Nuti, Giuseppe, Mahnoosh Mirghaemi, Philip Treleaven, and Chaiyakorn Yingsaeree, "Algorithmic Trading," *Computer*, Vol. 44, No. 11, 2011, pp. 61–69.

Ohm, Paul, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review,* Vol. 57, 2010, p. 1701.

Pasquale, Frank, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Boston: Harvard University Press, 2015.

Pearl, Judea, *Causality*, Cambridge: Cambridge University Press, 2009.

Roth, Alvin E., "The NRMP as a Labor Market: Understanding the Current Study of the Match," *Journal of the American Medical Association*, Vol. 275, No. 13, 1996, pp. 1054–1056.

Rudder, Christian, *Dataclysm: Love, Sex, Race, and Identity—What Our Online Lives Tell Us About Our Offline Selves*, New York: Crown Publishing, 2014.

Russell, Bertrand, *The Problems of Philosophy*, New York: Oxford University Press, [1912] 2001.

Salmon, Felix, "The Formula That Killed Wall Street," *Significance,* Vol. 9, No. 1, February 2012, pp. 16–20.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," paper presented to the Data and Discrimination: Converting Critical Concerns into Productive Inquiry preconference of the 64th Annual Meeting of the International Communication Association, Seattle, Wash., May 22, 2014.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot, "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature,* Vol. 529, No. 7587, January 28, 2016, pp. 484–489.

Sweeney, Latanya, "Discrimination in Online Ad Delivery," *ACM Queue,* Vol. 11, No. 3, April 2, 2013, p. 10.

Tett, Gillian, "Mapping Crime—Or Stirring Hate?" *Financial Times*, August 22, 2014. As of December 6, 2016:
https://www.ft.com/content/200bebee-28b9-11e4-8bda-00144feabdc0

Tufekci, Zeynep, "The Real Bias Built in at Facebook," *New York Times,* May 19, 2016. As of December 6, 2016:
http://www.nytimes.com/2016/05/19/opinion/the-real-bias-built-in-at-facebook.html

Turing, Alan M., "Computability and λ-Definability," *Journal of Symbolic Logic*, Vol. 2, No. 4, 1937a, pp. 153–163.

———, "On Computable Numbers, with an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society*, Vol. 2, No. 1, 1937b, pp. 230–265.

42 U.S. Code 3504, Discrimination in the Sale or Rental of Housing and Other Prohibited Practices, 1968.

42 U.S. Code 3505, Discrimination in Residential Real Estate–Related Transactions, 1968.

42 U.S. Code 3506, Discrimination in the Provision of Brokerage Services, 1968.

U.S. Department of Justice, Civil Rights Division, *Investigation of the Baltimore City Police Department,* Washington, D.C., August 10, 2016. As of December 6, 2016:
https://www.justice.gov/opa/file/883366/download

Valiant, Leslie, *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*, New York: Basic Books, 2013.

Wilson, James Q., and George L. Kelling, "Broken Windows: The Police and Neighborhood Safety," *Atlantic Monthly,* Vol. 249, No. 3, March 1982, pp. 29–38.

Ziewitz, Malte, "Governing Algorithms: Myth, Mess, and Methods," *Science, Technology, & Human Values,* Vol. 41, No. 1, September 30, 2016, pp. 3–16.

تؤثر الخوارزميات ووكلاء الذكاء الاصطناعي (أو "وكلاء الذكاء الاصطناعي" مجتمعة) على كثير من جوانب الحياة مثل: قراءة المقالات الإخبارية، والحصول على الائتمان، واستثمار رأس المال، من بين جملة أمور أخرى. وبسبب كفاءتها وسرعتها، تتخذ الخوارزميات القرارات وتنفذ الإجراءات نيابة عن البشر في هذه المجالات وكثير غيرها. وعلى الرغم من هذه المكاسب، هناك مخاوف بشأن المكننة السريعة للوظائف (حتى الوظائف المعرفية، مثل الصحافة والطب الإشعاعي). مع ذلك، لا يُظهر هذا الاتجاه أي علامات على التراجع. ومع استمرار ازدياد الاعتماد على وكلاء الذكاء الاصطناعي، ما العواقب والمخاطر المترتبة على هذا الاعتماد؟ ومن الضروري بلوغ فهم أفضل للمواقف تجاه الخوارزميات وأشكال التفاعل معها، ولا سيما بسبب هالة الموضوعية والتنزه عن الخطأ التي تضفيها ثقافة اليوم على الخوارزميات. ويوضح هذا التقرير بعض أوجه القصور في اتخاذ القرار الخوارزمي، ويحدد المواضيع الرئيسية التي تدور في فلك مشكلة الأخطاء والتحيزات الخوارزمية (مثل تغذية البيانات والتأثير المتباين الخوارزمي)، ويدرس بعض المقاربات لمكافحة هذه المشاكل. هذا التقرير على قدر كبير من الأهمية لصناع القرار والمنفذين الذين يسعون إلى اكتساب فهم أفضل لكيفية تأثير نشر الذكاء الاصطناعي على أصحاب الشأن لديهم. وهذا من شأنه التأثير على مجالات مثل العدالة الجنائية، والأشغال العامة، وإدارة الرعاية الاجتماعية.

$12.00

www.rand.org

ISBN-10 0-8330-9763-6
ISBN-13 978-0-8330-9763-7

Arabic Translation:
"An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence"
RR-1744/1-RC