Global Network
on Extremism & Technology

الذكاء الاصطناعي ومكافحة التطرف العنيف: كتاب تمهيدي

ماري شروتر

مشروع GNET من المشرعات الخاصة التي يقدمها المركز الدولي لدراسة الراديكالية، كينجز كوليدج لندن.

هذا التقرير بقلم ماري شروتر، زميلة مركاتور في التكنولوجيا الجديدة في العلاقات الدولية: إمكانات الذكاء الاصطناعي وحدوده لمنع التطرف العنيف على الإنترنت

الشبكة العالمية للتطرف والتكنولوجيا (GNET) مبادرة بحثية أكاديمية يدعمها منتدى الإنترنت العالمي لمكافحة الإرهاب (GIFCT)، وهي مستقلة ولكن تمولها الصناعة من أجل فهم أفضل لاستخدام الإرهابيين للتكنولوجيا والتصدي لهم. ويقوم المركز الدولي لدراسة الراديكالية (ICSR) بتنظيم والإشراف على فعاليات الشبكة العالمية للتطرف والتكنولوجيا (GNET)، بصفته مركزًا بحثيًا أكاديميًا داخل قسم دراسات الحروب في كينجز كوليدج لندن. والآراء والاستنتاجات الواردة في هذه الوثيقة آراءُ المؤلفين، ولا تُفسر على أنها تمثل آراء منتدى الإنترنت العالمي لمكافحة الإرهاب (GIFCT) ولا الشبكة العالمية للتطرف والتكنولوجيا (GNET) ولا المركز الدولي لدراسة الراديكالية (ICSR)، سواء كانت صريحة أو ضمنية.

ونود أن نشكر Tech Against Terrorism (التكنولوجيا في مواجهة الإرهاب) لما قدموا من دعم في هذا التقرير.

بيانات الاتصال

لأي أسئلة أو استفسارات، أو للحصول على نسخ أخرى من هذا التقرير، يرجى التواصل مع:

ICSR
King's College London
Strand
London WC2R 2LS
المملكة المتحدة

هاتف: **+44 20 7848 2098**
بريد إلكتروني: **mail@gnet-research.org**

تويتر: **@GNET_research**

الملخص التنفيذي

مثلما يحدث التطرف عبر الإنترنت، يقع على أرض الواقع أيضًا. ولكن لا يزال دور الإنترنت في ذلك محل خلاف. ولا شك، هناك مجتمعات راديكالية ومتطرفة على الإنترنت. ويبحث هذا التقرير في قدرة تطبيقات الذكاء الاصطناعي على المساهمة في مكافحة الراديكالية. ويحدد التقرير إمكانيات هذه التكنولوجيا بمختلف صورها، وحدودها، لدعم صانعي القرار والخبراء في شق غمار الضوضاء المحيطة بها، واتخاذ قرارات مستنيرة لا تتأثر بما حولها من ضجيج. وأكثر النتائج لفتًا للانتباه:

1. يمكن تعديل إعدادت محركات البحث وأنظمة التوصية للإشارة إلى المحتوى المعتدل ومكافحة الراديكالية

تمتاز محركات البحث وأنظمة التوصيات بإمكانيات كبيرة تمكنها من تأمين المساحات على الإنترنت وتقليل فرص مواجهة المحتوى المتطرف، ومن ثم منع التطرف العنيف. تساعدنا محركات البحث على التجول في خضم المعلومات التي يزخر بها الإنترنت، بما في ذلك المحتوى المتطرف. يمكن تعديل الخوارزميات لتشير إلى المحتوى المعتدل بدلًا من المحتوى المتطرف. وبالمثل، فإن أنظمة التوصية التي تقترح الفيديو أو الأغنية أو الفيلم التالي بناءً على سجل التصفح يمكنها أن تعزز وجهات النظر المتطرفة من خلال التوصية بالمحتوى التأكيدي. وقد يتصدى نظام التوصية المتوازن للروايات الخبيثة بمحتوى معارض أو ينشر معلومات عن مشاريع ونقاط اتصال تتعلق بمنع التطرف العنيف ومكافحته.

2. معالجة اللغات الطبيعية يمكنها مساعدتنا في ترجمة لغات الأقليات لتحسين إدارة المحتوى ودعم الإشراف على محتوى مواقع الويب المتخصصة على المدى البعيد

وتوفر معالجة اللغات الطبيعية (NLP) إمكانية إدارة المحتوى عبر الإنترنت، خاصة فيما يتعلق باللغات التي تتحدثها مجموعات صغيرة من الأشخاص فقط. وفيما يبدو غالبًا أن تعديل المحتوى بلغات الأقلية لا يدر أرباحًا تكفي للاستثمار فيها. وكثيرًا ما لا تتمتع المنصات الأصغر بالخبرة الفنية أو الموارد اللازمة لأنظمة إدارة المحتوى، بل يتطلب استخدام النماذج الحالية وقتًا وجهدًا كبيرين. وآخرون يدعمون ما يمكن اعتباره تطرفًا في تفسير معنى حرية التعبير، ومن ثم لا يريدون تقييد المستخدمين. وتساعدنا المعالجة اللغوية الطبيعية المحسنة في ترجمة المحتوى إلى لغات يعمل بها عدد كبير من المشرفين المدربين ذوي الخبرة. وتكتشف المعالجة اللغوية الطبيعية أنماطًا دلالية غير عادية على مواقع الويب أيضًا. وهذا قد يفيدنا في دعم اكتشاف الرسائل المهمة على المنصات التي لا تريد أو لا تستطيع الاستثمار في إدارة المحتوى. ومع ذلك، يجب أن تحترم هذه التدابير معايير الخصوصية وحقوق الإنسان في جميع الأوقات.

3. تفتقر معالجة المعلومات المضللة والتلاعب بالمحتوى عبر الإنترنت إلى الحلول الآلية

وحتى الآن لم تظهر أدوات آلية مقنعة ترصد المعلومات المضللة والتلاعب بالمحتوى وتعالجهما، وهذا قانوني بالرغم من أضراره. وهناك تحسن هائل في معرفة المستخدمين الرقمية يؤهلهم لفرض سيادتهم الرقمية، وفيما يبدو أن هذا النهج أنفع على المدى القصير.

4. ومهما كان الذكاء الاصطناعي خارقًا لن "يدق ناقوس الخطر" إذا ظهر شخص راديكالي على الإنترنت

ولا يستطيع الذكاء الاصطناعي العام أن يراقب المحتوى وسلوك الأفراد، بذكاء خارق، على الإنترنت و "يدق ناقوس الخطر" إذا اجتمعت مؤشرات الراديكالية معًا، وسيبقى من نسج الخيال لسببين. أولاً، لا توجد بيانات كافية لتغذية الخوارزميات بمعلومات محددة عن الراديكالية، ومتى يلوذ الراديكالي بالعنف. وما لم يكن هناك ابتكار تقني يسمح بإنشاء أنظمة موثوقة ببيانات أقل بكثير، فليس هناك ما يدعو للاستعانة بالمساعدة الفنية لأنها لا تستطيع التنبؤ بشئ موثوق بدون بيانات كافية عن الحالات السابقة. ولا تكفي معلوماتنا عن الراديكالية والإرهاب لإنشاء أي خوارزمية، بسبب ندرتهما وقلة تنوعهما لحسن الحظ. ثانيًا، يتطلب التنبؤ بسلوك الأفراد بيانات عن الأفراد يمكن تصنيفها بوضوح، وقد تتجرد عن الخصوصية جملة وتفصيلاً وقد تفضي إلى فرض المراقبة أكثر من ذي قبل. ولا يتوافق السيناريو الموضح أعلاه مع الديمقراطيات الليبرالية التي تتمتع في جوهرها بحقها في الخصوصية.

المحتويات

1 مقدمة

هناك اعتقاد شائع بأن الذكاء الاصطناعي سوف يُحدث ثورة في كل شيء، بما في ذلك الأمن القومي. ومازال السؤال عن تأثير الإنترنت في تسهيل الراديكالية بلا جواب، لكن الهجمات الإرهابية التي وقعت في هالي في ألمانيا الشرقية، وكريستشيرش في نيوزيلندا، وكنيس بواي في كاليفورنيا، مجرد ثلاثة أمثلة حديثة على أهمية دور الإنترنت في تعزيز الراديكالية اليوم.

كيف يساعدنا الذكاء الاصطناعي في مكافحة الراديكالية عبر الإنترنت؟ وتتفرع الخبرة في هذا الشأن إلى فروع عديدة، لكن نجدها في من يتمتعون بخلفيات في المجال الأمني ومكافحة الإرهاب من الباحثين والخبراء، وكذلك صانعي السياسات وخبراء التكنولوجيا الذين يعكفون على دراسة هذا المجال. ومشهد المعلومات الحالي يصعِّب على صانعي القرار تمييز الغث من السمين عند تصفية المعلومات في الوقت الراهن. ويلقي هذا التقرير الضوء على آخر التطورات في مجال الذكاء الاصطناعي، ويضعها في سياق جهود الديمقراطيات الليبرالية في مكافحة الراديكالية.

ومن باب التيسير عليهم، يلقي هذا المنشور الضوء على بعض حدود الذكاء الاصطناعي وإمكانياته في مكافحة الراديكالية عبر الإنترنت. ويقدم الفصل الثاني شرحًا موجزًا للمفاهيم والأفكار التي يقوم عليها الذكاء الاصطناعي. ونطرح في نهاية الفصل "رؤية متعمقة" ونولي اهتمامًا خاصًا لجودة البيانات والتلاعب في مجموعات البيانات وإخضاعها للأهواء. ويناقش الفصل الثالث إمكانات الابتكارات التكنولوجية القائمة على الذكاء الاصطناعي والقيود المفروضة عليها لتوفير مساحة "صحية" على الإنترنت تخلو من المحتوى الإرهابي والمواد الدعائية والمشاركة الزائفة. ونفترض أن هذه البيئة الصحية على الإنترنت تسهم في منع الراديكالية. ويقيِّم الفصل مجموعة من المفاهيم الشائعة القائمة على الذكاء الاصطناعي، بدءًا من التزييف العميق إلى جيوش البوتات التي تنشر أخبارًا زائفة، ويبين سبب قدرة محركات البحث وأنظمة التوصية، ولاسيما المعالجة اللغوية الطبيعية، على المساهمة في تحقيق هذا الهدف بطريقة أو بأخرى. وخصصنا الفصل الرابع "للذكاء الاصطناعي العام" الافتراضي، وهو نظام شامل يحدد الأفراد الخاضعين للراديكالية، ومن ثم يستطيع أن يساعد في إنفاذ القانون ومنع الجريمة قبل حدوثها. ويوضح هذا الفصل أيضًا أن تقنية الذكاء الاصطناعي لن يُكتب لها البقاء في المستقبل القريب إلا في عالم الخيال. وهذا يقودنا إلى مناقشة الأسباب الكامنة وراء ذلك الأمر. وفي الديمقراطيات الليبرالية، لا تُناقَش البيانات الضخمة، لا سيما ما يتعلق منها بالأمن التقليدي، إلا بعد حماية الخصوصية وإعطائها الأولوية. ونغوص في الفصل الرابع "غوصًا عميقًا" يشبع نهم الباحثين عن المزيد. ونختتم التقرير بالفصل الخامس.

يستند التقرير إلى مقابلات أعددنا لها بعض الترتيبات مع عدد من الباحثين وصانعي السياسات والمستشارين وممثلي القطاع الخاص. وتأثرت المواقف الواردة في هذا التقرير بم توصلت إليه نتائج البحوث المكتبية والمراقبة الإعلامية. وتحدثت إلى عدد من أصحاب المصلحة لأضع منظورًا متعدد التخصصات يراعي تشتت مشهد المعلومات من حولنا. وفُرضت على الأبحاث قيود واضحة لصلتها بمعلومات عن التعلم الآلي، وهذه المعلومات إما أنها لم تتحرر من قبضة الاستخبارات الأمنية أو قيود شركات القطاع الخاص.

2 ما هو الذكاء الاصطناعي؟

الذكاء الاصطناعي مصطلح شائع جدًا اليوم، ولكن ليس له تعريف موحد مشترك حول العالم. وهذا يرجع جزئيًا إلى أن دراسة الذكاء الاصطناعي موضوع سريع وشائع جدًا، ويطرح نتائج جديدة دائمًا ويطمس الحدود بين الحوسبة والإحصاءات والروبوتات. وبالرغم من عدم وجود إجماع على التعريف، فإنه يؤثر على معظم جوانب حياتنا. وهذا النوع من التكنولوجيا يحدد مشترواتنا التالية عبر الإنترنت، ويتحكم في يومياتنا، ويوجه سياراتنا بدون سائق.

ويوضح Alexa، مساعد أمازون الصوتي، مدى التطور الذي تحقق في أنظمة اتخاذ قراراتنا الآلية: ويستطيع Alexa أن ينظم لنا مواعيدنا بكل سهولة، بما في ذلك حجز تذاكر العروض الفنية، وحجز طاولة في مطعم، وطلب أوبر وإبلاغ أحبائنا بموعد وصولنا.[1] يشير مصطلح الذكاء الاصطناعي عمومًا إلى المجال الذي يعنى بدراسة الأنظمة التكنولوجية المؤتمتة والتكيفية. ويقوم الذكاء الاصطناعي بالمهام المطلوبة بدون توجيه مستمر، ويستطيع تحسين الأداء بالتعلم من التجارب السابقة.[2]

أُطلق اسم الذكاء الاصطناعي لأول مرة في مؤتمر عُقد في دارتموث كوليدج، هانوفر، نيوهامبشر في عام 1956. وحقق نجاحًا أوليًا دفع الباحثين إلى التفاؤل بأن استخدام الخوارزميات القائمة على الكمبيوتر سوف يحقق تقدمًا سريعًا. ونجحوا في هذه المراحل المبكرة في كتابة كود لحل المشاكل؛ وتضمنت البرامج عناصر معينة لتحسين الأداء بالتعلم. ودخل الذكاء الاصطناعي مرحلة "خريف العمر"، لضعف قدرات الذاكرات والمعالجات، ثم توقف الاستثمار في أبحاثه وقل الاهتمام بها في ستينيات القرن الماضي.

ولم يشهد الذكاء الاصطناعي هذه الضجة في القرن الحادي والعشرين إلا بفضل التطورات التقنية التي يقودها القطاع الخاص بشكل أساسي. وازدهر هذا المجال مع تزايد البرامج وحلول التخزين الشامل الرخيصة، وتعدد الخبراء المتخصصين وأصبح الوصول إلى البيانات أسهل. ما هي مزايا الذكاء الاصطناعي؟ أولاً، يسهل عملية تحليل البيانات المجمعة؛ لأنه أسرع في إجراء عملية التحليل وأكفأ من التقنيين. ثانيًا، هذه التكنولوجيا قادرة على العمل مع عدم اليقين، وهذه القدرة تمكنها من التنبؤ بالمستقبل. وهذه التنبؤات سواء كانت موثوقة أو غير موثوقة، لا يهم كثيرًا. إنها بالضبط قدرة الخوارزميات على التنبؤ، وهذا مصدر قوتها. وبالمقارنة، لا يستطيع العقل البشري اتخاذ القرارات بناءً على مجموعات البيانات الكبيرة والظروف المتنوعة وعدم اليقين. ويمكن اعتبار قوة التنبؤ على أنها قدرة مميزة للخوارزميات.

والتركيز على مصطلح "الذكاء الاصطناعي" في حد ذاته مضلل. فهو يشير إلى وجود تشابه مع الذكاء البشري أو عمليات التعلم البشري. وتُعد الأنظمة العصبية العميقة من تقنيات التعلم الآلي الخاصة التي ترتب فيها وحدات المعالجة في طبقات متعددة، وفكرتها مستوحاة بالفعل من بنية الدماغ البشري. ومع ذلك، تختلف قدرات هذه الأنظمة اختلافًا كبيرًا عن الخلايا العصبية البشرية. وتستطيع الخوارزمية ملء البيانات المفقودة، حتى في الحالات المعقدة، باستخدام نماذج التنبؤ، لكنها لا تستطيع أن تضفي على نتائجها معاني محددة. وتظهر أوجه الاختلاف بين ذكاء الإنسان والآلة بوضوح عند النظر إلى ما تستطيع الشبكة العصبية وما لا تستطيع تحقيقه. وعلى سبيل المثال، تستطيع الخوارزمية تحديد سرطان الثدي في مراحله المبكرة بدقة موثوقة تفوق قدرات الإنسان. لأن معدل خطئها في تحليل صور التصوير الشعاعي للثدي

1 Hao, K. (2019a), 'Inside Amazon's plan for Alexa to run your entire life', MIT Technology Review. متاح على: https://www.technologyreview.com/s/614676/amazon-alexa-will-run-your-life-data-privacy/

2 Reaktor & University of Helsinki (2018), 'How should we define AI?'. متاح على: https://course.elementsofai.com/1/1

أقل من معدل خطأ فني الأشعة.[3] من ناحية أخرى، لا تفهم الخوارزمية – ولا تستطيع تفسير معاني – مشاعر المريض ولا تتفاعل معها. فالتعاطف يحتاج إلى سنوات من الملاحظة، ناهيك عن الذكاء العاطفي الذي تفتقر إليه الخوارزميات. هذا فضلاً عن أن كلمة "ذكاء" في مصطلح الذكاء الاصطناعي تعني أن النظام قادر على توليد فكر أصلي، وهو أمر بعيد المنال حتمًا. يستطيع برنامج غوغل ألفا جو (AlphaGo) للذكاء الاصطناعي حساب التحركات الواعدة في لعبة Go شديدة التعقيد بسهولة، وهذا لا يعني أن AlphaGo يفهم اللعبة نفسها.[4] لا يستطيع AlphaGo تفسير سياق تلك التحركات، أو أنها بالنسبة له مجرد لعبة، أو ما جدوى هذه اللعبة أصلاً. ويمتاز الإنسان دون غيره من المخلوقات بقدرته على إسناد المعنى والسياق؛ لذا يعلم معظم الأطفال سبب استمتاعهم بممارسة الألعاب. ومع أن النظام يعجز عن تفسير سبب قيامه بما يفعل، فإنه يستطيع تحديد أفضل خطوة ممكنة في موقف معين وفقًا للهدف الذي يحتاج برنامجه إلى تحقيقه. ويحلل جميع الخيارات، ويحدد كيف يقلل مخاطرها حسابيًا، ثم يتعامل مع عدم اليقين.

وإذا دققنا النظر لوجدنا أن نقاط القوة والضعف في الذكاء الاصطناعي تنقسم إلى فئتين. وقد يُعتبر AlphaGo ذكاءً اصطناعيًا "ضيقًا" لأنه يؤدي مهمة واحدة، بينما يستطيع الذكاء الاصطناعي "العام" التعامل مع أي مهمة فكرية (ولا نجد هذا الذكاء الاصطناعي حاليًا إلا في الخيال العلمي). وذكاء النظام إما أن يكون ضعيفًا أو قويًا، وهذا ما نعبر عنه بقولنا ذكاء اصطناعي ضيق أو عام تباعًا. وينطبق مصطلح "الذكاء الاصطناعي الضيق" على أي نظام يتحلى بالذكاء بتحقيق النتائج المرجوة. وقد يكون الذكاء ضحلاً ويعتمد على هياكل خاطئة غالبًا: فمثلاً لا تستطيع الخوارزميات المدربة على تحديد القطارات في الصور الإشارة إلى القطار نفسه، ولكن يمكنها التعرف على المسارات المتوازية كثيرة التكرار التي تظهر على صور القطارات. واعتمدت الخوارزمية على هياكل خاطئة في شبكاتها العصبية لأنها قدمت النتيجة المرجوة.[5] وهذا الأمر مخاطره واضحة: لم تظهر تبعاته المحتملة بعد. ومن تبعاته المعروفة مثلاً أن أنظمة التعرف على الوجوه لا تستطيع التعرف على أصحاب البشرة الملونة بسهولة.[6] قد يكون للذكاء الاصطناعي العام عقل حقيقي أو وعي أو ذكاء فائق تستخدمه وسائل الإعلام السائدة في الإشارة إليه. ونكرر أن الأنظمة الفائقة الذكاء لا نجدها هكذا إلا في الخيال العلمي.

ولقد تغير معنى مصطلح الذكاء الاصطناعي بمرور الوقت. وفي الوقت الحاضر، مثلما يستخدم الذكاء الاصطناعي والتعلم الآلي بمعنى واحد في العديد من المنابر الإعلامية؛ يستخدمان هكذا في هذا التقرير أيضًا. وللتعلم الآلي نوعان شائعان عمومًا: أحدهما مراقَب والآخر غير مراقَب. ويعني التعلم الآلي المراقَب أن الخوارزمية تُدرب على تحليل البيانات بناءً على مجموعة بيانات تدريبية معينة تتضمن بيانات مسبقة التسمية. وهذه التسمية بمثابة قرار بين نوعين من البيانات "بيانات تناسب الحالة" أو "بيانات لا تناسب الحالة". ومن ثم فإن البيانات المسماة تعطي معنى لنقاط البيانات التي تتطلب مدخلات بشرية. ولنضرب مثلاً بصورة التفاحة. فإذا كانت كذلك فعلاً، فإنها تسمى "تفاحة". ويتطلب تدريب الخوارزمية الوصول إلى كمية كبيرة من البيانات الواضحة التسمية. ويمكننا استخدام مجموعة بيانات معينة لاختبار أداء الخوارزمية، ونجاحها في الاختبار يعني أنها تستطيع بعد ذلك تسمية البيانات الجديدة. وميزة التعلم المراقَب أن الخوارزمية تنظم البيانات حسب برمجتها تمامًا. وعملية تسمية البيانات يدويًا مرهقة ومكلفة. وكثيرون من مستخدمي الانترنت يسمون البيانات بأنفسهم، لأن بعض مواقع الويب تطرح أسئلة "أنا لست روبوتًا" المعروفة للتأمين. وقد تطلب تلك الأسئلة من المستخدم أن يعلِّم على جميع الصور التي يرى فيها سيارات مثلاً. وهنا تحديدًا يقوم المستخدم بتسمية البيانات. وهي طريقة للتحقق تطبقها غوغل في reCaptcha، وتستخدم نتائجها في ترويض مجموعات بيانات التعلم الآلي.[7] قد يُستفاد من البيانات المسماة في السيارات المسيرة بدون سائق مثلاً.

---

3 Hao, K. (2020), 'Google's AI breast cancer screening tool is learning to generalize across countries', MIT Technology Review. متاح على: https://www.technologyreview.com/f/615004/googles-ai-breast-cancer-screening-tool-is-learning-to-generalize-across-countries/

4 Gibney, E. (2017), 'Self-taught AI is best yet at strategy game Go', Nature. متاح على: https://www.nature.com/news/self-taught-ai-is-best-yet-at-strategy-game-go-1.22858

5 Thesing, L. et al. (2019), 'What do AI algorithms actually learn? – On false structures in deep learning', Arxiv. متاح على: https://arxiv.org/abs/1906.01478

6 Simonite, T. (2019), 'The best Algorithms Struggle to Recognize Black Faces Equally', Wired. متاح على: https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/

7 Google ReCaptcha (بلا تاريخ). متاح على: https://www.google.com/recaptcha/intro/v3.html

وهناك أيضًا التعلم الآلي غير المراقَب. وهنا يجب أن تعثر خوارزمية التعلم الآلي على أنماط في البيانات غير المسماة وغير المنظمة بدون ترويضها على التوافق بين بيانات المدخلات وبيانات المخرجات. وتعثر الخوارزميات غير المراقَبة على الأنماط داخل مجموعات البيانات دون أن تحتاج إلى تعليمات أو تصنيفات أخرى. ولا تعرف الخوارزمية مدى فائدة الأنماط التي تعثر عليها لمبرمجي الخوارزمية، وهنا يكمن التحدي. وقد تصل إلى نتائج لا صلة لها بالموضوع أصلاً. ومع ذلك توفر علينا جهودًا كثيفة في تسمية البيانات وتستفيد من هذا الكم الهائل من البيانات غير الموسومة والمتاحة للجميع. ويمكن استخدام التعلم غير المراقَب كخطوة أولى قبل التعامل مع خوارزمية التعلم المراقَب.

وكثيرًا ما تتحدث التقارير والأخبار عن التعلم العميق. وهي تقنية خاصة من تقنيات التعلم الآلي، وتعتمد على ربط عدة وحدات معالجة بشبكة واحدة يتيح نطاقها إمكانية تحليل المشكلات الأكثر تعقيدًا. وبالمثل، كثيرًا ما تتصدر الشبكات العصبية عناوين الأخبار. وفكرتها مستوحاة من بنية الدماغ البشري، وتتيح إمكانية تخزين المعلومات ومعالجتها في وقت واحد. وكان لهذه الشبكات دور كبير في اختراق البيانات الكبيرة، لأنها تتيح إمكانية معالجة كميات ضخمة من البيانات.

الغوص العميق

البيانات – الجودة والتحيز والتلاعب

لا يكاد يمر اجتماع أو مؤتمر عن الذكاء الاصطناعي إلا ويُستخدم فيه شعار "البيانات نفط الغد". لكن هل هذا الكلام صحيح فعلاً؟ بالتأكيد، يحتاج التعلم الآلي إلى الكثير من البيانات، والحصول عليها ليس سهلاً، ولكنه باهظ الثمن. وأيضًا، كلما تحسنت الخوارزميات، ازدادت بها البيانات. ومع ذلك، لكل نقطة من نقاط البيانات قيمة تميزها عن غيرها، حيث تنطبق قاعدة تناقص المردود. وما تتعلمه الخوارزمية من البيانات المبكرة أكثر من ما تتعلمه من تكرارها ولو للمرة المليون. وبالمثل، تعمل الخوارزميات جيدًا لاسيما إذا كانت بياناتها التدريبية قد أعدَّتها للعديد من الاحتمالات – مجموعة بيانات تتضمن عناصر عادية وغير عادية. لذا، فالبيانات ند النفط من حيث ارتفاع أسعارها وزيادة الطلب عليها (العديد من نماذج الأعمال الجديدة تعتمد على البيانات في استخدام التعلم الآلي) وفي الوقت الحالي، لا تتوفر هذه السلعة إلا في إيدي قلة محدودة نسبيًا. وبينما تساهم كل قطرة نفط في إجمالي إنتاج النفط، يختلف تأثير نقاط البيانات على قيمة البيانات ككل.

والخوارزميات عرضة للتحيز، وهو تشويه منهجي للواقع من خلال توظيف عينات البيانات. وسوف يؤدي التحيز في المدخلات حتمًا إلى تحيز في الناتج، وكلما ازداد، تضاعف تأثيره. وتبين وجود تحيز في مجموعات البيانات الحالية عدة مرات، لاسيما ضد النساء وأصحاب البشرة الملونة. على سبيل المثال، اضطرت أمازون إلى إزالة أداة مواردها البشرية المؤتمتة، لأنها تميز ضد النساء. وصنفت الخوارزمية الرجال أفضل من النساء لأنها دُربت على مستندات تطبيقية من العقد الماضي. ولا يخفى علينا أن الذكور يهيمنون على قطاع التكنولوجيا إلى حد كبير، وهذا من التحيز الذي ظهر في قرارات الخوارزمية.[8]

والبيانات الجيدة غير المتحيزة أمر حتمي للخوارزمية. وأسست الأمم المتحدة، بعد مشاورات طويلة، منظمة Tech Against Terrorism، التي أنشأت منصة تحليلات المحتوى الإرهابي (TCAP) لإنشاء هذه المجموعة من البيانات لتكون مرجعًا يستعين به القطاع الخاص والأوساط الأكاديمية والمجتمع المدني. وكان الهدف من Tech Against Terrorism في الأصل إدراج محتوى القاعدة وتنظيم داعش، لكنها أعلنت لاحقًا أنها ستوسع نطاق TCAP لتشمل الإرهاب اليميني المتطرف أيضًا. وهناك مجالات أخرى تستحق الاهتمام أيضًا، مثل الإرهاب الذي تغذيه أيديولوجية تكره المرأة، كما رأينا في حركة Incel. ومن الواضح أن مجموعات البيانات هذه يجب أن تتوافق مع معايير أمان البيانات وأن تأخذ في الاعتبار الآثار الصحية العقلية المحتملة للمراجعين.

ومع ذلك، فالبيانات عرضة للتلاعب. ويصعب، على الخبراء غير التقنيين أكثر من غيرهم، اكتشاف مجموعة البيانات التي عُبث بها من الخارج. وأثبت باحثون صينيون أن البيانات لا تحتاج بالضرورة إلى الكثير من التغييرات للتلاعب بالخوارزمية. وأسفرت تجاربهم عن اختيار سيارة ذاتية القيادة السير على الجانب العكسي من الطريق. وهذا يوضح هشاشة الأنظمة.[9] فيما يخص تطبيقات الذكاء الاصطناعي المصممة لمنع الراديكالية عبر الإنترنت، يمكن أن يؤدي التلاعب إلى نتيجة عكسية. ولنا أن نتخيل، على سبيل المثال، نظامًا لتعديل المحتوى الخادع يقرر إزالة ملفات تعريف الحزب السياسي المعارض في الفترة التي تسبق الانتخابات. ومن نقاط الضعف الأخرى التي تظهر في التدريب العدائي لأنظمة التعلم الآلي: التنافس بين نظاميْن لتحسين الأداء المتبادل. وعلى سبيل المثال، يقوم أحد أنظمة التعلم الآلي بإنشاء وجوه زائفة والآخر يصفيها ويستخلصها من مجموعة الوجوه الحقيقية. وكلما تطور أداء نظام التصفية، ازداد نظام تزييف الوجوه مهارة. ولم تتأكد حتى الآن العواقب المترتبة على ذلك.

---

8 Dastin, J. (2018), 'Amazon scraps secret AI recruiting tool that showed bias against women', Reuters Technology News. متاح على: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

9 Knight, W. (2019), 'Military artificial intelligence can be easily and dangerously fooled', MIT Technology Review. متاح على: https://www.technologyreview.com/2019/10/21/132277/military-artificial-intelligence-can-be-easily-and-dangerously-fooled/

3 الذكاء الاصطناعي في مواجهة الراديكالية عبر الإنترنت – ليس كل ما يلمع ذهبًا

كيف تستخدم الإحصاءات واتخاذ القرار الآلي في مكافحة الراديكالية عبر الإنترنت؟ كالعادة، ليس كل ما يلمع ذهبًا، وحول الذكاء الاصطناعي ضوضاء تصم الآذان. وهذا يُصعِّب على الخبراء غير التقنيين فهم مقدار المادة الموجودة بالفعل. ويتعمق هذا الفصل في طرح إمكانيات وقيود الابتكارات التكنولوجية الشائعة القائمة على التعلم الآلي. ويركز على العناصر السائدة في أحاديث العامة. من التزييف العميق إلى تعديل المحتوى الآلي ومحركات البحث ومعالجة اللغة الطبيعية، يهدف هذا الفصل إلى مساعدتنا في تقييم استخدامات التكنولوجيا في مكافحة الراديكالية. وهناك تداخل حتمي بين مختلف عناصر هذه البيئة، لاسيما أنها تتغير سريعًا وتتوالى تطوراتها بكل جديد.

3-1 تشكيل تجربة المستخدم عبر الإنترنت – الواضح الذي لا تخطئه العين

تتأثر تجربة المستخدم على الإنترنت بالتعلم الآلي تأثرًا ملحوظًا لأنه يشكل ما لا تخطئه أعين المستخدمين بكل سهولة. تتسم الخوارزميات بتنوع صورها وقدرتها الكبيرة على مواجهة الراديكالية، لأنها تسهم في تنقية الإنترنت من المحتوى الضار. ويتناول هذا التقرير محركات البحث وأنظمة التوصية والإدارة المؤتمتة للمحتوى.

تساعد **محركات البحث** المستخدمين في نهاية الأمر على التجول في ملايين مواقع الويب والعثور على المحتوى ذي الصلة عبر الإنترنت. وتقودنا محركات البحث إلى الاتجاه الصحيح، ضمن كتلة المعلومات والبيانات عبر الإنترنت؛ كأنها دليل من أدلة أرقام الهواتف في القرن الحادي والعشرين. وتُعد خوارزميات محركات البحث هي العمود الفقري للنجاح. وكان مشهد محركات البحث قبل عشر سنوات، أكثر تنوعًا، وفي نهاية الأمر، أصبحت الريادة لغوغل بفضل وصفتها السحرية، بعد أن عزز الثقة في أدواتها من خلال تقديم محتوى ذي صلة. وتلبي اليوم مليارات من طلبات البحث يوميًا، منها 15٪ استفسارات جديدة. ولا تعثر الخوارزميات سهلة الاستخدام على المعلومات المطلوبة فحسب، بل تتعرف أيضًا على الأخطاء الإملائية وتقترح تلقائيًا الكلمة التالية في شريط البحث. وفي النهاية، تحدد برمجة الخوارزمية المعلومات التي يجب تقديمها. وحسب ما ذُكر أن إمكانية الوصول إلى كتيبات القنابل على الإنترنت أدت مباشرة إلى أنشطة إرهابية، كما يتضح من حالة نويل فيلينتزاس وآسيا صدِّيقي اللتيْن استخدمتا، كما لاحظ وكيل مكتب التحقيقات الفيدرالي، مجلة تنظيم القاعدة Inspire، منشورات مدونة حول المتفجرات المحلية الصنع و كتاب The Anarchist Cookbook لصنع المتفجرات المحلية الصنع.[10] نفذت بريطانيا عملية كب كيك، وقام فيها جهاز الاستخبارات البريطاني (MI6) ومكاتب الاتصالات البريطانية (GCHQ) بإزالة دليل لصنع القنابل المحلية الصنع من مجلة Inspire واستبداله بوصفات "أفضل كب كيك في أمريكا". باختصار: من الأهمية القصوى بمكان أن نعرف سواءً كان العثور على أدلة صنع القنابل المؤقتة ممكنًا أو كان هناك تلاعب بالخوارزميات لتعجيز الباحثين عن المحتوى المتطرف على الإنترنت. وهذا لن يثني عباقرة التكنولوجيا تمامًا، ولكنه يصعِّب الوصول إلى ذلك المحتوى.

10 United States District Court, Eastern District of New York (2015), United States of America vs. Noelle Velentzas and Asia Siddiqui. Complaint and affidavit in support of arrest warrant, 2014R00196. متاح على: https://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/04/02/velentzas-siddiqui-complaint.pdf

أنظمة التوصية أداة ملائمة للعثور على المقطع أو الأغنية أو المقالة أو عنصر التسوق التالي بناءً على العناصر التي سبق استخدامها أو شراؤها. وتؤدي أنظمة التوصية إلى اكتشاف الأغنية الجديدة أو تسهيل العثور على الفراش المناسب لمرتبة اشتُريت حديثًا. ومع ذلك، فإن الخوارزميات التي تقترح أشياءً تاليةً يتعين النظر أو الاستماع إليها يمكنها أيضًا إنشاء **فقاعات تصفية** تعزز الافتراضات باقتراح مواد مماثلة. وهذا قد يؤدي أيضًا إلى تعزيز المواقف المتطرفة. ولا يستطيع المستهلك أن يختار كيفية التوصية بالأشياء أو تزكيتها بحرية، لعدم وضوح الكيفية التي تقترح بها الخوارزميات عناصرَ جديدةً على وسائل التواصل الاجتماعي ومواقع الموسيقى أو الفيديو أو الأفلام: المزيد من نفس الشيء، وجهات النظر المتعارضة أو كلاهما معًا. وتبين من البحث عن المحتوى المقترح تلقائيًا على منصات التواصل الاجتماعي الأكبر من عام 2019 أن خوارزميات YouTube على وجه الخصوص ساهمت في تعزيز وجهات النظر المتطرفة. وبمجرد مشاهدة مقطع فيديو لمحتوى متطرف أو هامشي، يوصَى بمحتوى مماثل.[11] هذا الأمر في حد ذاته مقلق لأن موقع YouTube هو أشهر مواقع التواصل الاجتماعي بين البالغين في الولايات المتحدة الأمريكية، وليس مستبعدًا أن يتلقى العديد من المستخدمين أخبارهم من الموقع.[12]

تعرضت وسائل التواصل الاجتماعي المعروفة لانتقادات شديدة مرارًا وتكرارًا لعدم تحركها بحسم كافٍ لمناهضة استغلال الإرهابيين منصاتها. وانطلقت دعوات تنادي بمسؤولية الوسطاء عن التعامل مع هذا المحتوى، نظرًا لأن وسائل التواصل الاجتماعي بمثابة منابر شبه عامة يقصدها الناس ويتبادلوا فيها أحاديثهم ويقوموا بأعمالهم. وهناك **منافذ خدمية متخصصة** لا حصر لها تقدم اليوم مشهدًا متنوعًا من الخدمات الإلكترونية عبر الإنترنت بكامل نطاقه وبنيته التحتية، بدءًا من تطبيقات التواصل التي تحافظ بدرجة كبيرة على عدم كشف الهوية، والمنصات المتخصصة التي ليس لديها إمكانات أو ميل لمتابعة المحتوى، إلى خدمات الاستضافة التي تسمح بنشر البيانات ومقاطع الفيديو الحية. وتعرضت مواقع الويب والخدمات المتعلقة بصناعة الألعاب مؤخرًا لانتقادات شديدة لعدم منع الاستخدام الضار.[13] توصلت جامعة سوانسيا في بحوث أجرتها مؤخرًا إلى كيفية استخدام تنظيم داعش مجموعة من خدمات الاستضافة لصحيفته الإلكترونية Rumiyah، ما أدى إلى إلغاء مركزية المحتوى وتعزيز التحديات لإزالته بسرعة وفعالية.[14] هناك حاجة ماسة إلى مزيد من البحوث والبيانات الوصفية حول كيفية استخدام الإرهابيين الخدمات المتخصصة.

2-3 إدارة المحتوى الذي أنشأه المستخدم

ثورة ويب 2.0 في التفاعل عبر الإنترنت. أتاح إمكانية الانتقال بعيدًا عن مواقع الويب الثابتة إلى تفاعلات تحدث في وقتها الفعلي من قبل أعداد كبيرة من المستخدمين في جميع أنحاء العالم. مع أن هذا الترابط العالمي يعزز العديد من المجتمعات المدنية ويدعمها على نطاق غير مسبوق، فقد فرض تحديات جديدة أيضًا على جهود مكافحة الراديكالية.

الإدارة المؤتمتة للمحتوى على منصات التواصل الاجتماعي تهدف إلى منع انتشار المحتوى الإرهابي. وتقوم خوارزميات التعلم الآلي بتصفية 98٪ من المحتوى الضار بالفعل على فيسبوك، كما ورد في أحدث تقرير لتقارير التقييم الذاتي للاتحاد الأوروبي حول ممارسة التضليل الإعلامي.[15] النسبة المتبقية 2٪ يؤشر عليها المستخدمون. وذكر موقع تويتر أنه يتصدى لعشرة حسابات في الثانية؛[16] ويزيل جوجل، مالك موقع يوتيوب، 80٪ من مقاطع الفيديو غير اللائقة قبل أن يشاهدها أحد، حسب ما ورد

11 Reed et al (2019), 'Radical Filter Bubbles', in the 2019 GRNTT Series, an Anthology, RUSI, London.

12 Perrin, A. & Anderson, M. (2019), 'Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018', Pew Research Centre. متاح على: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

13 Schlegel, L. (2020), 'Points, Ratings and Raiding the Sorcerer's Dungeon: Top-Down and Bottom-Up Gamification of Radicalisation and Extremist Violence'. متاح على: https://gnet-research.org/2020/02/17/points-rankings-raiding-the-sorcerers-dungeon-top-down-and-bottom-up-gamification-of-radicalization-and-extremist-violence/

14 Macdonald, S. et al (2019), 'Daesh, Twitter and the Social Media Ecosystem', The RUSI Journal, vol. 164, no. 4, pp.60–72.

15 فيسبوك (2019)، تقرير فيسبوك عن تنفيذ مدونة ممارسات التضليل الإعلامي. متاح على: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62681

16 Twitter (2019), Twitter Progress Report: مدونة قواعد الممارسة لمناهضة التضليل الإعلامي. متاح على: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62682

عنهما.[17] فيما يبدو أن هذا من مظاهر النجاح في الإدارة، ومن الإنصاف أن نقول إن إدارة المحتوى قد تحسنت خلال السنوات الأخيرة. ومع ذلك، لا تزال هناك ثغرات هائلة، لاسيما عند مغادرة مناطق اللغة القياسية. وفي الوقت الحالي، اقتصر مستودع فيسبوك اللغوي للتحقق من المعلومات على 14 لغة رسمية في أوروبا من أصل 26. وتخضع 15 دولة أفريقية للمراقبة الآن، بفضل التعاقد مع طرف ثالث، لكن هذا العدد أقل من ثلث دول القارة.[18] لم يتضح بعد إن كان هذا ينطبق على اللغات الرسمية فقط أو يشمل اللهجات أيضًا. ويعد حذف المحتوى المقدم بلغات الأقليات للمراجعة ظاهرة معروفة من أقاليم ودول أخرى متنوعة وكثيفة السكان، مثل الهند.[19]

وكما ناقشنا سابقًا، لا تستطيع الخوارزميات أن تضفي معنى على البيانات، ما يعني أن الخوارزميات تعجز عن فهم سياق حدوث السلوك الخبيث. وهناك أمثلة من سريلانكا توضح أن خوارزميات الفيسبوك عجزت عن تقييم السياق الثقافي متعدد الطبقات. ولقد تسللت المنشورات، قبل تفجيرات عيد الفصح في كولومبو في أبريل 2019، لأن الخوارزميات عجزت عن فهم مدى تعقد خطاب الكراهية فيها. وبالرغم من الجهود المبذولة للإبلاغ عن خطاب الكراهية، والتي أججت استقطاب المشاعر المعادية للمسلمين، فشل فيسبوك في إزالة هذا المحتوى أو توفير الإدارة المناسبة للمحتوى استجابة لهذا الأمر.[20] لتصنيف اللغة العامية المستخدمة في الحض على الكراهية، كان ينبغي أن تفهم خوارزميات إدارة المحتوى أعراق طرفي الحديث. وتتفاقم المشكلة: اللغات التي لا تستخدم الأبجدية اللاتينية كثيرًا ما "تُترجَم" إليها من باب التيسير. وبعض اللغات ليس بها صيغة زمنية نحوية تعبر عن المستقبل صراحةً، فكيف يبدو التهديد الذي يشير إلى المستقبل؟ إذا كنا نعتزم توسيع نطاق التصفية الآلية، فيجب أن نواجه إخفاقات هذا التصميم أولاً.

تحرص العديد من شركات التواصل الاجتماعي على منع الجهات الخبيثة من استغلال منصاتها. وتعتمد فعالية الخوارزميات في كشف المواد الدعائية أو النشاط الإرهابي أيضًا على جودة البيانات التي دُرِّبت الخوارزمية عليها ومدى توافرها. واعترف موقع فيسبوك بأن نقص بيانات التدريب كان سبب إخفاقه في تحديد وتصفية البث المباشر لعمليات إطلاق النار، مثل هجوم كرايستشيرش الذي بُث على الهواء مباشرة. ويستخدم الآن لقطات التقطتها كاميرات ضباط الشرطة البريطانية المحمولة أثناء تدريبات الإرهابيين.[21]

وتُعد **المعالجة اللغوية الطبيعية** (NLP) القائمة على التعلم الآلي من الابتكارات التي تساعدنا في إدارة المحتوى الذي أنشأه المستخدم وتوسيع نطاق الإشراف على المحتوى. وتعبر عن الإجراءات والأدوات الفنية لتحليل اللغة ومعالجتها. يتشعب تطبيق المعالجة اللغوية الطبيعية (NLP): تنتشر المعالجة اللغوية الطبيعية (NLP) في كل مكان، من بوتات الدردشة لدعم العملاء، وبرامج الإملاء، والترجمة الآلية إلى التحدث إلى Siri. وإذا أردنا أن نرى مدى التقدم الذي حققته التكنولوجيا في السنوات الأخيرة، فلننظر إلى ترجمة اللغات، على وجه الخصوص. وفي الماضي، كان نجاح الترجمة التي تُجرى عبر الإنترنت جزافيًا، أما الآن، باتت الترجمة تجرى بدقة يُعتمد عليها. ومع ذلك، لا تخلو الترجمة الآلية من العيوب ولا تؤخذ عن المترجمين الفوريين على علاتها. وظهرت خدمة الترجمة من غوغل (Google Translate) وتناقل المترجمون التحريريون أخبارها عندما ظهرت لقطة لترجمة "أنا ذكي" و "أنا جميل" من الإنجليزية إلى الإسبانية والفرنسية بصيغة المذكر. واستخدمت صيغة المؤنث في جملة "أنا جميلة ولست ذكية".[22] وإصلاح هذه العيوب قيد البحث والتطوير. وتوصلت شركة Baidu الصينية إلى تقنية جديدة تسمى التورية (Masking)، وتسمح لبرنامج الترجمة أن يتحرر من قيود

17 Google (2019), EC EU Code of Practice on Disinformation. متاح على: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62680

18 Africa Times (2019), 'Facebook expands fact-checking to 15 African nations'. متاح على: https://africatimes.com/2019/10/10/facebook-expands-fact-checking-to-15-african-nations/

19 Perrigo, B. (2019), 'Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch', Time. متاح على: https://time.com/5739688/facebook-hate-speech-languages/

20 Wijeratne, Y. (2019a), 'The Social Media Block isn't helping Sri Lanka', Slate. متاح على: https://slate.com/technology/2019/04/sri-lanka-social-media-block-disinformation.html
و Wijeratne, Y. (2019b), 'Big Tech is as Monolingual as Americans', Foreign Policy. متاح على: https://foreignpolicy.com/2019/05/07/big-tech-is-as-monolingual-as-americans/

21 Manthorpe, R. (2019), 'Police share "shooting" video with Facebook to help identify live-streamed attacks', SkyNews. متاح على: https://news.sky.com/story/police-share-shooting-video-with-facebook-to-help-identify-live-streamed-attacks-11843511

22 'Marta Ziosi', LinkedIn، متاح على: https://www.linkedin.com/posts/marta-ziosi-3342007a_googletranslate-googletranslate-women-activity-6603598322009808896-MQJX

ترجمة الكلمة بكلمة (الترجمة الحرفية) ويراعي السياق، ليحقق نتائج أدق.[23] وقد يُستفاد من هذا الأمر في التقارير الأخيرة حول حركة بوغالو اليمينية المتطرفة، والتي يُقال إنها تستخدم لغة مشفرة عبر الإنترنت للإفلات من عمليات الإزالة الآلية على منصات وسائل التواصل الاجتماعي.[24]

تمتاز التكنولوجيا بإمكانيات كبيرة لدعم إدارة المحتوى على مواقع الويب مع تفهم تام لحرية التعبير. وظهرت 8chan و 4chan و Gab، وهي من أشهر الأمثلة الدالة على الإيديولوجية اليمينية المتطرفة والعديد من صور الكراهية الأخرى المرتبطة بالجماعات مثل معاداة السامية وكراهية الأجانب وتفوق البيض. وقد يعزز نهج "انعدام السياسات" وجود البيئات المتطرفة التي يُسمح فيها بمشاركة كل شيء، باستثناء المحتوى غير القانوني مثل المواد الإباحية للأطفال، بموجب تشريعات الولايات المتحدة. والعجيب أن مطلقي النار في كنيس بواي، و ولمارت في إل باسو، والمسجد في كرايستشيرش في عام 2019، قد نشروا جميعًا في 8chan قبل ارتكابهم هجماتهم الإرهابية. ومن الضروري إجراء المزيد من البحوث الدقيقة، إلا أن هذه المنشورات الأخيرة تظهر بوضوح وسط المزاح والسخرية والكلمات المسيئة للغاية كما تعودناها على تلك المواقع. ويشيرون جميعًا إلى هجمات أخرى ويتبادلون رابطًا إلى بيان أو بث مباشر أو منشورات أخرى؛ ويقول مطلقو النار أنهم قد يموتون. وتكتسي المنشورات بنغمة رقيقة حنونة. ومن المتصور أن المنشور إذا استوفى ثمة مؤشرات، فإن المعالجة اللغوية الطبيعية يمكنها تحديد مستويات التهديد المتصاعدة. وقد تحتاج هذه المؤشرات إلى تعديل حسب خصائص المنصة.

وتستطيع المعالجة اللغوية الطبيعية (NLP)، عن طريق تحسين إدارة المحتوى، أن تخلق مرونة في مجتمعات الإنترنت باستخدام لغات الأقليات. وتخفق الإدارة المؤتمتة للمحتوى في تحقيق نتائج موثوقة للغات الأقليات. وبدلاً من تمني تحسن آداء الخوارزميات قريبًا بالرغم من عدم قدرة الحافز الاقتصادي على إجبار الشركات على الاستثمار في تجسينها، ربما نجد الحل في المعالجة اللغوية الطبيعية (NLP)، وقد يكمن الحل في تحسين الترجمة إلى لغات المشرفين المتمرسين والمدربين. وقد تراقب إدارة المحتوى لغات أخرى لا تحظى بتغطية جيدة إذا كان مستوى الترجمة الآلية مقبولاً.[25] مع ذلك، يجب أن تحترم التطبيقات المحتملة دائمًا معايير الخصوصية وتتوافق مع حقوق الإنسان.

3-3 محتوى مصطنع يوجهه الذكاء الاصطناعي – كيف تقلب الطاولة

ويؤدي التلاعب بالمحتوى إلى تسرب الفكر الراديكالي إلى الخطاب العام، ويسهل الراديكالية ويحرك العنف في دنيا الواقع. لم يعُد التضليل السياسي استراتيجية جديدة، لكن إمكانية الوصول إلى جماهير كبيرة بمعدلات غير مسبوقة بضغطة زر لتوجيه النقاشات العامة تطرح تحديات جديدة. ويبحث هذا الفصل في طرق مواجهة المتصيدين والبوتات والأخبار الزائفة والتزييف العميق.

المتصيدون أو **البوتات** عبارة عن حسابات تواصل اجتماعي تنشر محتوى معينًا أو تخلق تفاعلاً مصطنعًا على منصات التواصل الاجتماعي. "يمكن برمجة هذه البوتات لأداء مهام مرتبطة عادةً بالتفاعل البشري، بما في ذلك متابعة المستخدمين، وتفضيل التغريدات، والرسائل المباشرة (DM) لمستخدمين آخرين، والأهم من ذلك، يمكنها تغريد المحتوى، وإعادة تغريد أي شيء نشرته مجموعة مستخدمين أو تتميز بوسم محدد."[26] لا تحتاج برمجة البوت معرفة تقنية متطورة، ويمكن القيام بها بسهولة، بمساعدة كتيبات متاحة بسهولة عبر الإنترنت.[27]

23 Baidu Research (2019), 'Baidu's Pre-training Model ERNIE Achieves New NLP Benchmark Record'. متاح على: http://research.baidu.com/Blog/index-view?id=128

24 Owen, T. (2020), 'The Boogaloo Bois are all over Facebook', Vice. متاح على: https://www.vice.com/en_us/article/7kpm4x/the-boogaloo-bois-are-all-over-facebook

25 Wijeratne, Y. (2019b).

26 Symantec Security Response (2018), 'How to Spot a Twitter Bot', Symantec Blogs/Election Security. متاح على: https://www.symantec.com/blogs/election-security/spot-twitter-bot

27 Agarwal, A. (2017), 'How to write a Twitter Bot in 5 Minutes', Digital Inspiration. متاح على: https://www.labnol.org/internet/write-twitter-bot/27902/

يُطلق على المتصيدين المستخدمين بأعداد كبيرة اسم شبكة أو جيش المتصيدين أو البوتات. ويؤثر المحتوى المُتلاعب به الذي تُنسقه تطبيقات التراسل المتعددة على المواقف العامة أو الخطاب العام وفقًا لأجندة الفرد ذاتها. ومن أشهر الأمثلة على ذلك التدخل الروسي في الانتخابات الأمريكية لعام 2016، حيث دعمت الاشتباكات الوهمية المفروضة استراتيجيًا حملة دونالد ترامب وهاجمت هيلاري كلينتون. وتشير التقديرات إلى أن 5٪ إلى 15٪ من الحسابات على الإنترنت زائفة (هذه الأرقام محل خلاف).[28] وبحسب دراسة أجرتها Pew Research، فإن أكثر خمسمائة بوت نشاطًا على تويتر مسؤولة عن 22٪ من الروابط المغرَّدة، فيما يمثل أكثر خمسمائة شخص نشاطًا حوالي 6٪ فقط. وفي الوقت نفسه، فإن 66٪ من الحسابات التي تتبادل روابط المواقع الأكثر شعبية بوتات و 34٪ فقط من البشر.[29] أجرت كاترزينا بروتشكيفيش مؤخرًا بحثًا استقصائيًا أوضحت فيه الحاجة إلى تنظيم مزرعة التروللت،[30] وقد سبق لها العمل لدى شركة تروللت بولندية لمدة ستة أشهر. وكانت تدير هي وزملاؤها محادثات عبر الإنترنت لصالح العملاء بمقابل، ومن بينهم جهات البث العامة. ولم يتضح بعد مدى ترجمة المشاركة عبر الإنترنت إلى أصوات فعلية،[31] لكن لا يقبل السياسيون ولا المؤسسات في الدول الديمقراطية كسب الحجج بالمال.

يرى مارك زكربيرغ، الرئيس التنفيذي لشركة فيسبوك، أن الذكاء الاصطناعي هو الحل لإدارة المحتوى، ويتضمن تحديد المشاركات الوهمية وإزالتها أيًا كان نوعها. وحسب زكربيرغ، تستطيع الأنظمة الآلية فقط معالجة محتوى مليون مستخدم بلغات مختلفة ومن خلفيات ثقافية متنوعة.[32] ومع ذلك، لم تتضح التفاصيل بعد. واعترف زكربيرغ أيضًا في جلسة استماع في مجلس الشيوخ الأمريكي في 2018 أن الذكاء الاصطناعي قد يكون جاهزًا في غضون خمس إلى عشر سنوات لاكتشاف الفروق الدقيقة في اللغة، لكن لم تظهر التطورات التقنية بعد.[33] تدعي الحلول التكنولوجية الحالية أن تحديد البوتات ممكن. وهناك افتراض أن البوت الذي يُعد لغرض معين يستطيع أن ينشئ موضوعًا واحدًا أو محتوى موضوعيًا ضيقًا للغاية ويشارك فيهما، على عكس الإنسان لأنه قد يهتم بعدد من الموضوعات أكبر. وتشمل معلومات التحليل الإضافية تاريخ ووقت إنشاء الحساب.[34] تتعارض هذه التكنولوجيا الواعدة مع النتائج التي توصل إليها مركز الناتو للتميز في ريغا. أثبتت التحريات التي أجريت مؤخرًا، وشملت فيسبوك وتويتر وإنستغرام ويوتيوب، أن تحديد المشاركة المصطنعة وإزالتها غير كافٍ.[35] وتمكن الباحثون، مقابل 300 يورو فقط، من شراء 3,530 تعليقًا و 25,750 إعجابًا و 20,000 مشاهدة و 5,100 متابع. وأخفقت المنصات في تصنيف السلوك أو الحسابات غير الأصلية: بعد أربعة أسابيع من الشراء، كانت أربعة عناصر من كل خمسة عناصر لا تزال على الإنترنت. وحتى بعد الإبلاغ عن إحدى العينات، ظل 95٪ من المحتوى على الإنترنت بعد مرور ثلاثة أسابيع على إخطار مواقع الويب. وحيث أن الجهات الفاعلة مصممة على نشر المحتوى الضار عبر حسابات زائفة أو جيوش البوتات، يجب أن يكون لدى المنصات نهج استباقي لتحديد الحسابات الزائفة حتى لا تصبح إدارة المحتوى كالحرث في الماء.

الأخبار الكاذبة أو **الأخبار التافهة** تطرح محتوى مختلقًا أو معلومات خاطئة أو نظريات مؤامرة بشكل مباشر، ولا تتعارض مع القانون بالضرورة ولكنها تمثل ضررًا بالتأكيد. وهناك مصطلحات أشد تمييزًا تفرق بين المعلومات المضللة والمعلومات الخاطئة. أما الأولى فهناك من يتعمد نشرها، والثانية تتفرق عن غير قصد. وكلٌ منهما يساعد على تسلل الفكر المتطرف إلى الخطاب العام، ما يسهل الراديكالية ويحرك العنف

28 Burns, J. (2018), 'How many Social Media Users are Real People?', Gizmodo. متاح على: https://gizmodo.com/how-many-social-media-users-are-real-people-1826447042

29 Wojcik, S. et al. (2017), 'Bots in the Twittersphere', Pew Research Centre. متاح على: https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/

30 Davies, C. (2019), 'Undercover reporter reveals life in a Polish troll farm', The Guardian. متاح على: https://www.theguardian.com/world/2019/nov/01/undercover-reporter-reveals-life-in-a-polish-troll-farm

31 Eckert, S. et al. (2019), 'Die Like Fabrik', Sueddeutsche Zeitung. متاح على: https://www.sueddeutsche.de/digital/paidlikes-gekaufte-likes-facebook-instagram-youtube-1.4728833

32 Cao, S. (2019), 'Facebook's AI Chief Explains How Algorithms Are Policing Content – And Whether It Works', The Observer. متاح على: https://observer.com/2019/12/facebook-artificial-intelligence-chief-explain-content-moderation-policy-limitation/

33 Harwell, D. (2018), 'AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how', The Washington Post. متاح على: https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/

34 Gupta, S. (2017), 'A Quick Guide to Identify Twitterbots Using AI', Hackernoon. متاح على: https://hackernoon.com/a-quick-guide-to-identify-twitterbots-using-ai-c3dc3a7b817f

35 Bay, S. & Fredheim R. (2019), 'Falling Behind: How social media companies are failing to combat inauthentic behaviour online', NATO STRATCOM COE. متاح على: https://www.stratcomcoe.org/how-social-media-companies-are-failing-combat-inauthentic-behaviour-online

في دنيا الواقع. وقد تكون المعلومات المضللة جزءًا من استراتيجية سياسية، وإذا انتشرت انتشارًا فعالاً (ربما عبر الحسابات الزائفة المذكورة أعلاه وجيوش البوتات)، تأثر بها الخطاب العام. فقد كانت "بيتزاغيت"، مثلاً، نظرية مؤامرة من الحملة الانتخابية الأمريكية عام 2016. وبعد تسريب رسائل بريد جون بوديستا الإلكتروني، وكان حينئذ مدير حملة المرشحة الرئاسية الديمقراطية هيلاري كلينتون، ادعى المعارضون أن بريده الإلكتروني به رسائل مشفرة تربط بين مسؤولين كبار في الحزب الديمقراطي وبين الاتجار بالبشر وجنس الأطفال. وخلال الحملة الانتخابية، نشر أنصار اليمين المتطرف هذه النظرية على مواقع الصور 4chan و 8chan وكذلك على Reddit وتويتر. وقيل أن بعض المطاعم سهلت جرائم مشتهي الأطفال المزعومين. وتلقى أصحاب هذه المطاعم والعاملون فيها تهديدات بالقتل وغيره. وفي النهاية، قرر إدغار ماديسون ويلش، متأثرًا بالمنشورات عبر الإنترنت، أن يذهب إلى أحد هذه المطاعم. وأطلق ثلاث طلقات. ولم يصب أحد. وبعد حادث إطلاق النار، أنكر في التحقيقات وأن هذه المعلومات مجرد أخبار كاذبة.

ومع أن المحتوى لا يمثل انتهاكًا للقانون، يعاقب أصحاب هذه المنصات انتهاكات المعايير العامة التي وضعها المجتمع نفسه. وقد تتعارض تصفية الأخبار الزائفة مع نموذج الأعمال المتبع في شركات التواصل الاجتماعي. ويزيد المحتوى الجذاب من تجمع المستخدمين ومشاركتهم، ويشجع الناس على قضاء المزيد من أوقاتهم على مواقع الويب، ما يمكّن الشركات من تجميع المزيد من البيانات عن عملائها. وهذه البيانات هي العمود الفقري لنموذجها المالي، لأنها تحسّن الإعلان الموجه وتزيد عوائده. ومع ذلك، تبذل مختلف الجهات الفاعلة جهودًا للتعامل مع الأخبار الزائفة. ولقد طوَّر باحثون من جامعة واترلو الكندية أداةً قائمةً على الذكاء الاصطناعي لدعم التحقق من صحة المعلومات بمستوى غير مسبوق. وبمجرد تحقق النظام من بيانات مقال ما مقابل مصادر أخرى، يقدم مؤشرًا عن الأخبار سواء كانت زائفة أو لا. ويرى الباحثون أن النظام قد حالفه التوفيق تسع مرات من أصل عشرة.[36] وقد تكون هذه الخطوة مهمة في مكافحة الأخبار الزائفة.

وتطرح أداة Newsguard، وهي جهد مشكور للقطاع الخاص من مايكروسوفت عملاق التكنولوجيا، مثالاً على عدم كفاءته. و Newsguard عبارة عن وظيفة إضافية لمتصفحات الويب تقوم بقياس مصداقية المنافذ الإعلامية في صورة تقييمات. وتشير إلى مصداقية المعلومات بوسم صغير يظهر على وسائل التواصل الاجتماعي. وهو حل مُرهق: يُطلب من المستخدم بشكل استباقي تنزيل هذه الوظيفة الإضافية، ولكنها لا تساعد بعد ذلك في تقييم مقالات معينة وإنما تقدم تصنيفًا عامًا للمنفذ الإعلامي فحسب. وما كان لهذه الأداة أن تساعد في فضيحة بيتزاغيت المذكورة أعلاه، لأنها انتشرت من خلال الحسابات الخاصة. وصُنفت Breitbart (بريتبارت)، وهي وسيلة لنشر تفوق البيض والأيديولوجية اليمينية المتطرفة، أو قناة الدعاية الروسية RT بوسم أخضر عام، أما Newsguard فتسلط الضوء في النص أدناه على وسمها أن هذه المواقع عليها "قيود كبيرة". وهاجمت Breitbart فيسبوك أيضًا: وأطلق فيسبوك قسمًا "إخباريًا" يعرض قصصًا إخباريةً من منافذ إعلامية تم التحقق منها. وتم تحديد هذه المنافذ الإعلامية بالتعاون مع الصحفيين، وتلتزم بتوجيهات فيسبوك الداخلية لمناهضة خطاب الكراهية والمحتوى المتصيد. وأثار انضمام Breitbart إلى المنافذ الإعلامية احتجاجات. ويدافع فيسبوك عن قراره حتى الآن، متذرعًا بحرية التعبير. وفي تلك الأثناء، أعلن تويتر، في أعقاب الانتخابات العامة البريطانية عام 2019، فرض حظر على جميع الإعلانات السياسية.[37]

يؤدي حظر المحتوى المتطرف عمومًا إلى فرضية توفير مساحة أنقى على الإنترنت عن طريق الحد من فرص الدخول في مواجهة مع المحتوى المتطرف على الإنترنت. ومع ذلك، فإنها تشكل دائمًا جانبًا من الرد على المحتوى الضار على الإنترنت، لأنها لا تنظر في الأسباب الجذرية للرأي المعبر عنه.

---

36 Grant, M. (2019), 'New tool uses AI to flag fake news for media fact-checkers', Waterloo News. متاح على: https://uwaterloo.ca/news/news/new-tool-uses-ai-flag-fake-news-media-fact-checkers

37 لا تزال التفرقة بين الإعلان السياسي وإعلان القضايا محل خلاف، وليس لهما تعريفات تحظى بقبول عالمي. وأفضت الصعوبات الناشئة عن ذلك إلى انطلاق دعوات للتعامل مع جميع الإعلانات بمعيار واحد صارم لتعزيز الشفافية والسماح بدراسة تأثيرها. للمزيد من المعلومات: Frederik J. Zuiderveen Borgesius et al. (2018): Online Political Microtargeting: Promises and Threats for Democracy. Utrecht Law Review, 14 (1). 82-96. متاح على: https://www.ivir.nl/publicaties/download/UtrechtLawReview.pdf; و Call Universal Advertising Transparency by default (2020). متاح على: https://epd.eu/wp-content/uploads/2020/09/joint-call-for-universal-ads-transparency.pdf

التزييف العميق عبارة عن نسخة متطرفة من بيانات ملفقة تم التلاعب بها، ويعني في جوهره "التقوُّل على شخص بشئ لم يقله". تتيح أحدث التطورات التكنولوجية للمستخدمين إنشاء مقاطع فيديو لأشخاص معينين باستخدام تعبيرات وجوههم وأصواتهم. فنجد مقاطع فيديو مسجلة لسياسيين يتكلمون، تبدو واقعية للغاية، مع أنهم في الحقيقة لم ينطقوا بكلمة واحدة مما جاء فيها. فقد حاول سياسي هندي في حملته الانتخابية الأخيرة أن يخطب ود شريحة سكانية متعددة اللغات؛ فاستخدم مقطع فيديو مزيفًا تزييفًا عميقًا، ما أثار ردود فعل متباينة.[38] وحذرت بعض المنظمات من تقنية التزييف العميق، وعممت النتائج التي توصلت إليها، مثل الفيديو الذي أيد فيه بوريس جونسون وخصمه جيريمي كوربين كلٌ منهما الآخر في انتخابات ديسمبر 2019. ونجد غالبية البيانات الملفقة في المواد الإباحية، ما يجعل النساء أكثر ضحايا هذه التكنولوجيا الجديدة. وهذا يعني أن امرأة قد تجد مقاطع فيديو إباحية من بطولتها دون علمها أو موافقتها. وفيما يبدو أن هذه التقنية، إذا استُخدمت مع أدوات **إخفاء الهوية**، قد تلحق ضررًا بالغًا. وتعمل هذه الأدوات على تغيير الصور ومقاطع الفيديو بطريقة يستحيل بها على الخوارزميات أن تكشف زيف الإصدار الجديد المعدَّل تعديلاً طفيفًا باعتباره الوجه الأصلي. وإنما تتعامل معه كعنصر جديد تمامًا. ومع ذلك، يستطيع المستخدمون التعرف على الوجه الأصلي في النسخة الجديدة المعدَّلة. وهذا قد يصعِّب إزالتها بطريقة سريعة وفعالة. وأنشأ منتدى الإنترنت العالمي لمكافحة الإرهاب (GIFCT) الذي تقوده الصناعة اتحادًا لتبادل الهاشات – قاعدة بيانات "بصمات" المحتوى الضار الرقمية، وتسمى هاشات.[39] ويسعى المنتدى إلى تعزيز فعاليتها بالتعاون بين شركات مختلفة.[40] لم يتضح بعد مدى قدرة قاعدة البيانات على الصمود للاستخدام المنهجي لبرمجيات إخفاء الهوية، لاسيما أن المحتوى المتطرف منتشر انتشارًا استراتيجيًا من خلال مجموعة من الجهات الفاعلة وعبر المنصات.

يحتاج التزييف الواقعي العميق إلى معرفة متخصصة – لاسيما إذا كان الهدف منها خداع الناس. وتقف المعرفة التقنية اللازمة حائلاً دون التوسع السريع في هذه التكنولوجيا، لاسيما إذا كان تحقيق الهدف ذاته ممكنًا بطرق أخرى أسهل. وأيضًا، يرى هوانغ أن أدوات إخفاء الهوية قد تُعرض صاحبها للانكشاف والانفضاح، وفي هذا خطر قائم. وتتراجع إغراءات أدوات التزييف العميق في التأثير على اتجاه الحملات بفضل سياسات الحظر وخشية الانفضاح بين العامة.[41] استخدمت تويتر وسومه الجديدة على المحتوى المتلاعَب به لأول مرة على المحتوى الذي أنشأه مدير وسائل التواصل الاجتماعي بالبيت الأبيض.[42] تنص السياسة على أنها ستبلغ عن مقاطع الفيديو أو الصور المُتلاعب بها ولكن لن تزيلها، ما لم يكن هذا المحتوى يهدد سلامة الشخص المعني.

تتطلب السيطرة على المعلومات المضللة والمشاركة المصطنعة معرفة رقمية أقوى بين المستخدمين. وتوضح الأبحاث التي تُجرى حول نشر الأخبار الزائفة على تويتر أن الأكاذيب تنتشر بسرعة أكبر وعلى نطاق أوسع من الحقيقة، وهذا بسبب التفاعل البشري. وتعزز البوتات رواجها، لكنها لا تسبب انتشار الأكاذيب على نطاق واسع. وأرجع الباحثون سبب الاستجابة العاطفية والجدة النسبية للمحتوى إلى هذا الانتشار.[43] تظهر النتائج بوضوح عدم وجود بديل عن التعليم المناسب لتمكين المستخدمين من التجول عبر الإنترنت بخفة ومرونة.

38 Christopher, N. (2020), 'We've just seen the First Use of Deepfakes in an Indian Election Campaign', Vice. متاح على: https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

39 'GIFCT, 'Joint Tech Innovation. متاح على: https://www.gifct.org/joint-tech-innovation/

40 Llansó, E. (2019), 'Platforms want centralised Censorship. That should scare you', Wired. متاح على: https://www.wired.com/story/platforms-centralized-censorship/; and Windwehr, S. and York, Jillian (2020), 'One Database to rule them all: The invisible Content Cartel that undermines the freedom of expression online', EFF. متاح على: https://www.eff.org/deeplinks/2020/08/one-database-rule-them-all-invisible-content-cartel-undermines-freedom-1.

41 Hwang, T. (2020), 'Deepfakes – A grounded threat assesment', Centre for Security and Emerging Technology.

42 Dent, S. (2020), 'Twitter labels video retweeted by Trump as "manipulated data"', Engadget Online. متاح على: https://www.engadget.com/2020/03/09/twitter-labels-trump-retweet-manipulated-media/

43 Vosoughi, S. et al. (2018), 'The spread of true and false news online', Science vol. 359, no. 6380, pp.1146–51. متاح على: https://science.sciencemag.org/content/359/6380/1146

4 التنبؤ بالراديكالية قبل حدوثها – الذكاء الاصطناعي العام لإنفاذ القانون

من السهل أن نتخيل الغرفة التي يحدث فيها السحر؛ دعنا نستعير هذه الصورة من Minority Report الكلاسيكي الخيالي العلمي: شاشة زرقاء كبيرة تعمل باللمس، تعرض نتائج آلة فائقة الذكاء يُفترض أنها تساعد في إنفاذ القانون. وتعتمد النتيجة المعروضة على البيانات المتاحة عن الأفراد وكذلك السلوك المباشر عبر الإنترنت. تنطلق صفارات إنذار النظام التحذيرية بمجرد رصد تزايد عوامل الخطر البالغ وتشير إلى ارتفاع خطير في مستوى الراديكالية. ووفقًا للسلوك المرصود، سوف يرسل النظام حينئذٍ الوحدات المناسبة إلى الموقع. وهذا النظام يمكّن الشرطة من التحرك السريع قبل حدوث أي شيء، بفضل القدرة التنبؤية للنظام القوي القائم على الذكاء الاصطناعي. وقد يبدو هذا السيناريو مغريًا، حتى لو كان مبالغًا فيه ويذكرنا بالخيال العلمي أكثر منه بالواقع. وهناك من يتلذذ بهذه التطورات عند الربط بين التقنيات الجديدة والأمن. يركز هذا الفصل فقط على أسطورة الذكاء الاصطناعي العام الخارق لمراقبة المحتوى وسلوك الأفراد على الإنترنت لمواجهة الراديكالية.

وتعكف مشاريع الشرطة التنبؤية على استكشاف كيف يستطيع الذكاء الاصطناعي أن يساعد جهات إنفاذ القانون في عملها. وهذه المشاريع عبارة عن تطبيقات للتعلم الآلي التي تتنبأ بالجرائم قبل وقوعها بناءً على علاقات إحصائية معينة لدعم إنفاذ القانون.[44] وتُعد فعالية هذه الأنظمة مثار جدل قوي. وعلى سبيل المثال، توقفت شرطة كنت في المملكة المتحدة عن استخدام البرنامج الأمريكي PredPol للتنبؤ بالجرائم، لأن قيمتها المضافة لم تكن مقنعة.[45] أفادت منظمة الحقوق المدنية Big Brother Watch بأن الأقليات تخضع لتدابير بوليسية مفرطة تُعيد تأجيج التحيز ضد مناطق معينة في مشاريع الشرطة التنبؤية، وتؤدي زيادة الدوريات في المناطق التي ثبت تاريخيًا أنها أكثر عرضة للجرائم إلى تزايد الإبلاغ عن الجرائم وتنشيء حلقة محكمة لتعزيز التحيز الهيكلي.[46] وينطوي هذا التحول نحو استخدام مؤشرات التنبؤ بالجريمة على تفكير تفسيري وصور غير سببية من التفكير في المخاطر. ويشير التغيير إلى تحرك نحو التأكيد على أهمية السياق في تحليل المخاطر ويمكن اعتباره خطوةً بعيدًا عن التنميط، الذي يُنظر إليه على أنه غير عادل وتمييزي في العديد من المجتمعات.[47] أكد التنميط العنصري أو الإثني على العلاقة بين إنفاذ القانون والمجتمعات التعددية.[48]

وقد يبدو النهج القائم على المؤشرات مغريًا في جهود مكافحة الراديكالية عبر الإنترنت. وقد تُتخذ قائمة المؤشرات المستندة إلى السلوك عبر الإنترنت وكذلك المحتوى المستهلك أساسًا لدعم البحث عن الأفراد الراديكاليين حاليًا. ومع ذلك، يصعب تخيل الانتقال من أنظمة الشرطة التنبؤية الحالية إلى أعمال مكافحة الراديكالية، لأسبابٍ ثلاثةٍ في المقام الأول.

والعقبة الأولى هي عدم وضوح أو فهم العمليات الراديكالية بدقة.[49] ولم يتضح بعد متى يتحرك الفرد المتطرف تحديدًا لارتكاب جريمة، ما يبرر التدخل. ولحسن الحظ، لا تحدث

---

44 Moses, B. L. & Chan, J. (2018), 'Algorithmic prediction in policing: assumptions, evaluation, and accountability', Policing and Society, vol. 28, no. 7, pp.806–22.

45 Big Brother Watch Submission to the Centre for Data Ethics and Innovations (2019), 'Bias in Algorithmic Decision Making (Crime and Justice)'. متاح على: https://bigbrotherwatch.org.uk/wp-content/uploads/2019/06/Big-Brother-Watch-submission-to-the-Centre-for-Data-Ethics-and-Innovation-Bias-in-Algorithmic-Decision-Making-Crime-and-Justice-June-2019.pdf

46 المرجع نفسه.

47 Monaghan, J. & Molnar, A. (2016), 'Radicalisation theories, policing practices, and "the future of terrorism?"', Critical Studies on Terrorism, vol. 9, no. 3, pp.393–413.

48 Open Society Foundations (2019), 'Ethnic Profiling: What it is and Why it must end'، متاح على: https://www.opensocietyfoundations.org/explainers/ethnic-profiling-what-it-and-why-it-must-end

49 انظر Monaghan و Molnar.

الراديكالية والإرهاب بقدرٍ كافٍ لتوفير مجموعة من البيانات الموثوقة. والراديكالية عملية فردية معقدة للغاية، ومع أن الباحثين قد حددوا عناصر معينة يتكرر حدوثها في عمليات الراديكالية،[50] فإن المعلومات المتاحة لا تكفي لتدريب الخوارزمية. وتحتاج أنظمة الذكاء الاصطناعي الحالية إلى قدر هائل من البيانات لتطوير قدراتها التنبؤية. ولا يبدو الأمر واعدًا، ما لم تحدث طفرة تكنولوجية في تكنولوجيا الذكاء الاصطناعي تمكِّننا من التعامل مع البيانات المحدودة جدًا. ولا نرى اليوم دلائل تبشر بهذا التحول.

تعمل الأنظمة الشُرطية التنبؤية الحالية بناءً على افتراضات جماعية إلى جانب معلومات تحدد مواقع المناطق المعرضة للجريمة ووقت وقوعها. وهذا يعني أن هناك مزيجًا من المعلومات المفتوحة المصدر والبيانات الحكومية والبيانات التي تقدمها الشركات الخاصة يغذي الخوارزميات لتتمكَّن من طرح تنبؤات مستنيرة. وباتت الافتراضات الأساسية أكثر اقتصاديةً وأقرب إلى الاختيار العقلاني. ولنضرب مثلاً بعملية سطو: بمجرد نجاح المجرم في ارتكاب جريمة ما في وقت ومكان ما، من المرجح أن يرتكب جريمته التالية في موقع ووقت مشابهين لينجح مجددًا. وتكمن الفكرة في أن المجرم يريد الحد من المخاطر مع تحقيق أقصى قدر ممكن من النجاح. ولا تسري هذه الافتراضات بالضرورة على الراديكالية والإرهاب. وهذا يعني أن اختيار الإرهاب يخضع لمنطق عقلاني، ولكنه لا يحدث بالطريقة ذاتها كما في جريمة السطو. وعلى النقيض من ذلك، لايزال الموت في سبيل الهدف السامي عامل جذب واضح في حملة داعش الدعائية "الموت واحد – فليكن شهادة" للانضمام إلى الخلافة.[51]

والسبب الثالث هو القيود التي تفرضها الديمقراطيات الليبرالية النابعة من فكرة أن الفرد ينعم بحماية الدولة ومن الدولة. وكتجربة فكرية: ما المطلوب للتنبؤ بالسلوك الفردي؟ سوف تحتاج الخوارزمية التي تتنبأ بسلوك الفرد إلى بيانات أكثر مختلفة عن المعلومات المتاحة عن الجماعة. أي أنها تستلزم بيانات عن سلوك الأفراد غير مجهولة المصدر، وكلما كثُرت كان أفضل، لضمان موثوقية التنبؤات. وهذا يتطلب مراقبة السلوك الفردي بطريقة غير مسبوقة: من الضروري تغطية المجتمع بأكمله لحظة بلحظة. وهذا لا يتوافق مع حقوق الخصوصية الحالية. ولا تقبله المجتمعات الحرة لدواعٍ أخلاقية ومعنوية. وقد لا تُفلت من تداعيته الحقوقُ الأساسية، مثل حرية التعبير والصحافة وتكوين الجمعيات وسرية الاتصالات، وما إلى ذلك.[52] ولن يزيدنا هذا في الأساس إلا مرارة وألمًا.

50 انظر Neumann.
51 Kingsley, P. (2014), 'Who is behind Isis's terrifying online propaganda operation?', The Guardian. متاح على: https://www.theguardian.com/world/2014/jun/23/who-behind-isis-propaganda-operation-iraq
52 Ganor, B. (2019), 'Artifical or Human: A New Era of Counterterrorism Intelligence?', Studies in Conflict and Terrorism.

الغوص العميق

الذكاء الاصطناعي الديمقراطي عن طريق التصميم

تعمل الخوارزميات على البيانات وتعطُّشها للمزيد من البيانات لا يرتوي أبدًا. ويجب أن تُدرَّب أولاً على مجموعة ضخمة من البيانات، ثم تُختبر بمزيد من البيانات لمواصلة طريقها وشق غمار المزيد من المعلومات. ومن غير المستغرب أن تتبادر إلى الذهن حماية الخصوصية والبيانات، لاسيما مع قضايا الذكاء الاصطناعي والأمن.

وبدلاً من محاولة تنظيم أنظمة صنع القرار المؤتمتة لتستوفي معايير المجتمعات الديمقراطية، ينبغي نسج القيم الديمقراطية داخل تصميم التكنولوجيا في المقام الأول. وينبغي أن تكون "الخصوصية" هي الوضع الافتراضي للتطوير التكنولوجي، بمعنى معالجة بيانات المستخدم وفقًا لأعلى معايير الخصوصية، ما لم يوافق المستخدمون على الإفصاح عن بياناتهم. وتظهر تأثيراته على تنظيم المحتوى على منصات وسائل التواصل الاجتماعي والتتبع الجماعي للبيانات السلوكية الشخصية. وعلاوة على ذلك، يتعين أن تقدم الأنظمة نتائج شفافة أو تفسيرات تمكِّن المشغلين البشريين من قياس تقييمات الخوارزمية وتحديد مدى مصداقية النتائج. ويُعد هذا بديلاً واضحًا لما يُعرف "بذكاء الصندوق الأسود الاصطناعي" الحالي، ولا يمكن تفسير نتائجه. وقد تؤدي زيادة الشفافية في مجال أنظمة التوصية أو تنظيم المحتوى، مثلاً، إلى تعزيز البحوث والمصلحة العامة، ما يؤدي إلى فهم الراديكالية عبر الإنترنت بصورة أفضل. ولا يمكن تطبيق المساءلة في عمليات صنع القرار إلا بتعزيز شفافية الذكاء الاصطناعي والثقة فيه. وتخضع تطبيقات صنع القرار المؤتمتة لعمليات تدقيق تضمن التنفيذ القانوني وتقدم حوافز تعزز مبدأ الإنصاف والديمقراطية في الذكاء الاصطناعي.

5 الخلاصة

وكان الهدف من هذا التقرير أن نناقش كيف تعزز التقنيات الداعمة للذكاء الاصطناعي جهود مكافحة الراديكالية عبر الإنترنت.

يقدم الذكاء الاصطناعي فرصًا جديدةً لتحليل البيانات الضخمة والتنبؤ بالمستقبل. وهناك مجال لتطبيق التكنولوجيا في حدود ضيقة تدعم منع التطرف العنيف وتقليل مخاطر مواجهة المحتوى الراديكالي عبر الإنترنت. ومن أهم الأدوات الواعدة القائمة على الذكاء الاصطناعي محركات البحث وأنظمة التوصية والمعالجة اللغوية الطبيعية (NLP). وتوفر المعالجة اللغوية الطبيعية (NLP) إمكانية إدارة المحتوى عبر الإنترنت، خاصة فيما يتعلق باللغات التي تتحدثها مجموعات صغيرة من الأشخاص فقط. والعوائد المالية المفترض أن تحققها المنصات الكبيرة لتستثمر في إدارة محتوى لغات الأقليات – ولاسيما المشرفين على المحتوى البشري – ليست كبيرة بما يكفي في الغالب. وكثيرًا ما لا تتمتع المنصات الأصغر بالخبرة الفنية أو الموارد اللازمة لأنظمة إدارة المحتوى، بل يتطلب استخدام النماذج الحالية وقتًا وجهدًا كبيرين. وهناك منصات أخرى تدعم ما يُعد تطرفًا في تفسير معنى حرية التعبير، بزعم أنها لا تريد تقييد المستخدمين. وتساعدنا المعالجة اللغوية الطبيعية (NLP) المحسنة في ترجمة المحتوى إلى لغات يجيدها المشرفون المدرَّبون. ويمكنها أيضًا اكتشاف أنماطٍ دلالية غير عادية على مواقع الويب التي تُعرض عن الاستثمار في إدارة المحتوى. ومع ذلك، يجب أن تحترم هذه التدابير معايير الخصوصية وحقوق الإنسان دائمًا.

ولا تزال إدارة المحتوى على منصات التواصل الاجتماعي الكبيرة تمثل تحديًا. وتقف الخوارزميات عاجزة عن تصفية المحتوى الضار في خضم هذا العدد الهائل من اللغات المستخدمة جنبًا إلى جنب مع نطاق ملوَّن من السياق الثقافي. نحتاج إلى خطاب عام وشامل حول ما يسمى "مواد المنطقة الرمادية" – المحتوى الضار لكنه قانوني. وينبغي أن يتوافق المجتمع كله على أرضية مشتركة تتضح فيها حدود حرية التعبير عبر الإنترنت. وينبغي ألا يُترك هذا القرار للشركات الخاصة وحدها. وبالمثل، لم يكتمل تطوير تقنية الذكاء الاصطناعي، ولا يمكنها محاربة المشاركة المصطنعة عبر الإنترنت، بما في ذلك المتصيدون والبوتات والأخبار الزائفة. ولا تزال هذه الأجهزة حرة طليقة لم تُكتشف بالكامل. ويجب إشراك المستخدمين في الأمر وتعليمهم الالتزام بسلوك مسؤول على الإنترنت يؤدي إلى السيادة الرقمية؛ وينبغي أن يسمح تصميم المنصات بوجود نظام "إخطار وإجراء" يتسم بالشفافية. ومع ذلك، يجب ألا نلقي المسؤولية على عاتق العملاء أو المستخدمين وحدهم. ويجب تعزيز الأمن والأمان على الإنترنت بسياسات تثني عن نشر المحتوى الضار أو الملفق عبر الإنترنت وتمنع نماذج الأعمال التي تعطي الأولوية للمحتوى الضار، لأنها تزيد من المشاركة عبر الإنترنت، وبالتالي تعزز مكاسب الإعلانات. والخلاصة أن يجتنب مشغلو المنصات إزالة الكثير من المحتوى والتعدي على حرية التعبير مع استمرار اتخاذ التدابير المناسبة لمنع المحتوى الضار.

وربما اتضح لنا الآن أن الذكاء الاصطناعي العام، وهو نظام فائق الذكاء، ليس خيارًا للتنبؤ براديكالية الأفراد عبر الإنترنت لسببين. الأول تقني: في ظل الوضع الحالي لتكنولوجيا الذكاء الاصطناعي، فإن الخوارزميات تحتاج إلى كميات هائلة من البيانات لطرح تنبؤات مفيدة عن المستقبل. ولحسن الحظ، لا تحدث الراديكالية والإرهاب بالقدر الكافي لإنتاج بيانات كافية لذكاء اصطناعي عام يتنبأ بسلوك الأفراد نحو الراديكالية عبر الإنترنت. وبلغت الإيجابيات الكاذبة والسلبيات الكاذبة معدلات لا تطاق. ومن الأجدى أن نستثمر في الموارد البشرية. ويتعلق السبب الثاني بالخصوصية: النظام الذي يراقب سلوك الأفراد عبر الإنترنت لحظة بلحظة، ويخزن البيانات ويحللها، لن يمتثل لمعايير الخصوصية في الديمقراطيات الليبرالية. وقد يفضي إلى مراقبة المجتمع بأسره.

وينبغي أن تُوضع معايير واضحة لتطبيق التقنيات القائمة على الذكاء الاصطناعي على المدى البعيد. وينبغي أن تحمي هذه المعايير المستخدمين من افتقار أدوات صنع القرار المؤتمتة إلى العدل والإنصاف، من خلال مجموعات البيانات المتحيزة أو المعالجة التمييزية للمحتوى مثلاً على أساس النوع أو الجنس أو الدين أو أي خصائص أخرى يكفل حمايتها قانونٌ حقوق الإنسان. ويجب أن تُعرض نتائج الخوارزميات بشفافية تسمح بالمساءلة عن القرارات المبنية على حسابات خوارزمية. ويجب أن تسلك التكنولوجيا طريق المستقبل بتطوير ذكاء اصطناعي يحترم حقوق الخصوصية ويخلو من التمييز، ويفهمه المشغل.

المشهد السياسي

كتب هذا القسم أرميدا فان ريج ولوسي توماس، وهما باحثتان مشاركتان في معهد السياسات في كينجز كوليدج لندن. ويلقي نظرة عامة على المشهد السياسي وعلاقته بهذا التقرير.

مقدمة

يمثل منع تمجيد العنف والإرهاب، وانتشار المعلومات المضللة، وغيرها من صور المحتوى المتطرف عبر الإنترنت تحديات تواجهها الجهات الفاعلة السياسية ومنصات التكنولوجيا في جميع أنحاء العالم. الذكاء الاصطناعي ومكافحة التطرف العنيف: كتاب تمهيدي، تقرير لمنتدى الإنترنت العالمي لمكافحة الإرهاب (GIFCT)، يلقي نظرة عامة شاملة على الفرص والتحديات التي يقدمها الذكاء الاصطناعي في مجال مكافحة التطرف العنيف.

ويتعرض صناع السياسات على الصعيدين الوطني والدولي، وكذلك شركات التكنولوجيا، لضغوط متزايدة للإسراع بإدارة المحتوى المتطرف وإزالته بفعالية أكبر. ومن الأسباب الداعية إلى ذلك وقوع أضرار فعلية ومأساوية، من حيث تكرارها واتساع نطاقها، نتيجة للمحتوى الضار عبر الإنترنت، مثل إطلاق النار على مسجد كرايستشيرش في مارس 2019، وإطلاق النار على كنيسة تشارلستون في الولايات المتحدة الأمريكية عام 2015، وإطلاق النار على مسجد مدينة كيبيك الكندية عام 2017. وبذلت صناعة التكنولوجيا وصانعو السياسات الوطنيين ومتعددي الجنسيات جهودًا حثيثة فيما بعد لإزالة بعض المحتوى المتطرف الذي ينتجه تنظيم الدولة الإسلامية والجماعات الجهادية العنيفة، فضلاً عن المحتوى الذي ينادي بالتعصب للبيض وكره النساء ومعاداة السامية وكراهية الإسلام.

الذكاء الاصطناعي ومكافحة التطرف العنيف: كتاب تمهيدي يتناول بالوصف والتحليل تقنيات الذكاء الاصطناعي الحالية التي صُممت لتعزيز وتعجيل والتدقيق في إدارة المحتوى على الإنترنت وإزالته. ومنها الأدوات "المدربة" على اللغات لتحديد المحتوى الضار والإبلاغ عنه، والتكنولوجيا التي تكتشف فيديوهات التزييف العميق، فضلاً عن عمليات التعلم الآلي لبناء أدوات خوارزمية لتحديد الهوية. ويستعرض التقرير العديد من التحديات والعقبات التي تحول دون الانتشار الفعال لهذه التقنيات. أولاً، تستطيع أنظمة التوصية القائمة على تقنية الذكاء الاصطناعي ذاتها أن تقود المستخدمين إلى "جحور" المحتوى الضار بشكل متزايد. ثانيًا، التركيز على إدارة محتوى اللغات الأوروبية يؤدي إلى إغفال المحتوى غير الظاهر وعدم الاعتراف بوجوده وعدم إدارته كما يجب. ثالثًا، لا توجد حاليًا أي تدابير فعالة لمواجهة المعلومات المضللة عبر الإنترنت، من وجهتيْ النظر التقنية والأخلاقية. وأخيرًا، يعتمد نظام الذكاء الاصطناعي العام الذي يراقب التواصل المباشر عبر الإنترنت على مستويات مستحيلة من البيانات المعروفة المصدر، ما قد يهدد الخصوصية وحقوق حرية التعبير.

وفي هذا التقرير، نستعرض طرق التعامل مع هذه الفرص والتحديات لدى تسع جهات رئيسية فاعلة في السياسات الوطنية والإقليمية: كندا وفرنسا واليابان وغانا ونيوزيلندا والمملكة المتحدة والولايات المتحدة الأمريكية والمفوضية الأوروبية والمديرية التنفيذية لمكافحة الإرهاب التابعة للأمم المتحدة. ونقدم لمحة عامة لهذه الجهود كل على حدة، ونختتم بتقديم بعض التوصيات السياسية.

الذكاء الاصطناعي ومكافحة التطرف العنيف: التعامل مع التحديات وتقييم التطورات الجديدة

كندا

طوَّرت الحكومة الكندية استراتيجية قوية لمكافحة الإرهاب، وتمثل جهودها ومبادراتها لمكافحة التطرف العنيف عبر الإنترنت جانبًا واحدًا من سياسة أعم وأشمل لمكافحة التطرف العنيف. وكغيرها من الحكومات، للأسف، كان استثمارها واهتمامها بمكافحة التطرف العنيف عبر الإنترنت نتيجةً لوقوع ضرر فعلي.

وفي أواخر يناير 2017، أطلق اكسندر بيسونيت، من سكان كيبيك، النار على مصلين في المركز الثقافي الإسلامي في مدينة كيبيك، ما أسفر عن مقتل ستة وإصابة خمسة. وتوصلت التحقيقات لاحقًا إلى أن بيسونيت، قبل إطلاق النار، كان عنصرًا نشطًا في الدوائر اليمينية المتطرفة والعنصرية على الإنترنت، وكان يتابع بانتظام حسابات أصحاب نظريات المؤامرة والقوميين البيض وشخصيات اليمين المتطرف على الإنترنت مثل بن شابيرو وأليكس جونز صاحب موقع InfoWars الإلكتروني.[53]

وعلى عكس مرتكبي العديد من الهجمات الإرهابية المعروفة الأخرى التي يغذيها المحتوى عبر الإنترنت، لم يصدر عن بيسونيت بيان أو تصريح بنواياه عبر الإنترنت.[54] ومع ذلك، باتت البيانات الإرهابية اتجاهًا عامًا متزايدًا يمكن مكافحته بالذكاء الاصطناعي. وكثيرًا ما تشير بيانات اليمين المتطرف بعضها إلى بعض؛ الإعجاب بالهجمات السابقة أو الأخيرة أو تكرار الميمات أو مختصرات الإنترنت مثلاً. ويستطيع الذكاء الاصطناعي أن يساعدنا في تحديد عدد مرات تحميل المحتوى الضار، مثل البيانات اليمينية المتطرفة، للتدخل قبل وقوع الهجمات على أرض الواقع.

تتعامل كندا مع التطرف العنيف عبر الإنترنت، كما أوضحت في استراتيجيتها الوطنية لمكافحة راديكالية العنف، بطريقة[55] ثلاثية الأبعاد: صياغة رسائل مضادة مع المجتمع المدني، ودعم أبحاث مكافحة التطرف العنيف ومشاركتها في المبادرات الدولية وتعزيز شراكتها مع شركات التكنولوجيا. والثالث، على وجه الخصوص، هو المساحة التي استثمرت فيها كندا في العلاقة بين الذكاء الاصطناعي–مكافحة التطرف العنيف.

الأهم من ذلك أن كندا، في عام 2019، كلفت مبادرة Tech Against Terrorism الدولية التي ترعاها الأمم المتحدة وتعمل مع صناعة التكنولوجيا العالمية بتطوير منصة تحليلات المحتوى الإرهابي (TCAP)،[56] وهي قاعدة بيانات محملة بمواد ومحتوى إرهابي تم التحقق منها مستقاة من مجموعات البيانات الحالية ومصادر مفتوحة. والمأمول أن تعمل منصة تحليلات المحتوى الإرهابي (TCAP) كخدمة تنبيه فوري للمحتوى الإرهابي والمتطرف العنيف على منصات الإنترنت الصغيرة: سيتم تبادل المحتوى الضار المؤكَّد على هذه المنصات بسرعة مع فرقها القائمة بإدارة المحتوى والتصرف بناءً عليه. وسوف تعمل منصة تحليلات المحتوى الإرهابي (TCAP)، على المدى المتوسط والبعيد، كأرشيف تاريخي للتحليل الأكاديمي الكمي والنوعي المحسَّن.[57]

---

53 Riga, A. (17 أبريل 2018)، 'Quebec Mosque Killer Confided He Wished He Had Shot More People, Court Told'، Montreal Gazette. متاح: https://montrealgazette.com/news/local-news/quebec-mosque-shooter-alexandre-bissonnette-trawled-trumps-twitter-feed/.
انظر أيضًا: Mahrouse, G. (2018), 'Minimizing and denying racial violence: Insights from the Quebec Mosque shooting', Canadian Journal of Women and the Law, vol. 30, no. 3, pp.471–93.

54 على سبيل المثال، روبرت باورز (مرتكب حادث إطلاق النار على كنيس في بيتسبرغ عام 2018)، ديلان روف (إطلاق النار على كنيسة تشارلستون عام 2015)، برينتون تارانت (إطلاق النار على مسجد كرايستشيرش عام 2019)، باتريك كروسيوس (إطلاق النار عام 2019 في إل باسو)، أندرز بريفيك (مذبحة أوتويا 2011)، ونشر كثيرون غيرهم بيانات عبر مختلف منصات الإنترنت قبل هجماتهم بوقت قصير. انظر: Ware, J. (2020), Testament to Murder: The Violent Far-Right's Increasing Use of Terrorist Manifestos – Policy Brief, International Centre for Counter-Terrorism – The Hague. متاح: https://icct.nl/publication/testament-to-murder-the-violent-far-rights-increasing-use-of-terrorist-manifestos/

55 National Strategy on Countering Radicalization to Violence', Public Safety Canada'. متاح: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-strtg-cntrng-rdclztn-vlnc/index-en.aspx#s7

56 تناولنا منصة تحليلات المحتوى الإرهابي (TCAP) أيضًا في قسم المشهد السياسي من تقرير الشبكة العالمية للتطرف والتكنولوجيا (GNET) عن "تحليل شفرة الكراهية: استخدام التحليل التجريبي للنصوص في تصنيف المحتوى الإرهابي." متاح: https://gnet-research.org/wp-content/uploads/2020/09/GNET-Report-Decoding-Hate-Using-Experimental-Text-Analysis-to-Classify-Terrorist-Content.pdf

57 'Press Release: Tech Against Terrorism Participates in UN General Assembly Week in New York', Tech Against Terrorism. متاح: https://www.techagainstterrorism.org/2019/10/08/press-release-tech-against-terrorism-participates-in-un-general-assembly-week-in-new-york/

من أهداف منصة تحليلات المحتوى الإرهابي (TCAP) المعلنة، في فضاء الذكاء الاصطناعي على وجه التحديد، دعم نظام إيكولوجي لمصنِّفات المحتوى الخوارزمية.[58] ويوضح التقرير التمهيدي للشبكة العالمية للتطرف والتكنولوجيا (GNET)، الذكاء الاصطناعي ومكافحة التطرف العنيف، أن "الخوارزميات تحتاج إلى كميات هائلة من البيانات لطرح تنبؤات مفيدة عن المستقبل".[59] تعتمد آليات إدارة المحتوى المؤتمتة المستندة إلى التعلم الآلي والمعالجة اللغوية الطبيعية على تحليل البيانات الضخمة لتدريب الذكاء الاصطناعي لفهم حقيقة البيانات، وتَعَلم التعرف على عناصر فيديوهات تنظيم داعش الدعائية (الشعارات والرايات وما إلى ذلك) لتحديد مقاطع الفيديو المقبلة وتمييزها بالعناصر ذاتها أو ما يماثلها. وتُعد منصة تحليلات المحتوى الإرهابي (TCAP)، أول منصة موحدة للمحتوى الإرهابي عبر الإنترنت، بمثابة منجم ذهب حقيقي لأنها تزخر ببيانات ومعلومات يحتاجها المطورون في تصميم خوارزميات التعلم الآلي للتعرف على المواد الإرهابية وتصنيفها.

وقد تمثل منصة تحليلات المحتوى الإرهابي (TCAP) قفزة تقنية كبيرة إلى الأمام في مكافحة التطرف العنيف عبر الإنترنت لأنها تزخر بمحتوى إرهابي مؤكَّد يُعد أرشيفًا تاريخيًا من مختلف منصات الإنترنت. وأوضحت الحكومة الكندية، بصفتها راعية مشاركة للمنصة، مدى قدرة الاستثمارات الموجهة والذكية في المبادرات عبر القطاعات على توفير فرص للتعاون بين الأوساط الأكاديمية والصناعية والمجتمع المدني في دفع الذكاء الاصطناعي قدمًا في سياق مكافحة التطرف العنيف.

المفوضية الأوروبية

ذكرت المفوضية الأوروبية في كتابها الأبيض بشأن الذكاء الاصطناعي الصادر في فبراير 2020، أن "أدوات الذكاء الاصطناعي يمكنها أن تقدم فرصة جيدة لتعزيز حماية مواطني الاتحاد الأوروبي من الجرائم والأعمال الإرهابية".[60] يريد الاتحاد الأوروبي اتباع نهج مزدوج لاستخدام الذكاء الاصطناعي: تنظيمي ويركز على الاستثمار. ويُعنى على وجه الخصوص بتمكين "الذكاء الاصطناعي الجدير بالثقة" عن طريق بناء إطار تنظيمي سليم للمساعدة في حماية المواطنين الأوروبيين وكذلك "إنشاء سوق داخلي غير احتكاكي" لتطوير الذكاء الاصطناعي.[61] "الذكاء الاصطناعي الجدير بالثقة، في هذه الحالة، يعني وجود تطبيقات دقيقة وقوية تقنيًا".[62] يعتزم الاتحاد الأوروبي أيضًا زيادة الاستثمار في الذكاء الاصطناعي إلى ما لا يقل عن 20 مليار يورو سنويًا بحلول عام 2030.[63]

عينت المفوضية الأوروبية فريق خبراء رفيع المستوى للذكاء الاصطناعي (AI HLEG) في عام 2019. ووضع هذا الفريق سبعة معايير لضمان موثوقية الذكاء الاصطناعي. وهذه المبادئ السبعة هي: الرقابة والوكالة البشرية؛ السلامة والمتانة التقنية؛ الخصوصية وحوكمة البيانات؛ الشفافية؛ التنوع وعدم التمييز والإنصاف؛ الرفاه المجتمعي والبيئي؛ والمساءلة.[64] بناءً على هذا، يدعو كتاب المفوضية الأبيض إلى نظام إيكولوجي من الثقة يضمن حماية الحقوق الأساسية.[65]

وتباينت شركات التكنولوجيا الكبرى في استجابتها للكتاب الأبيض المعني بالذكاء الاصطناعي. ودعت غوغل الاتحاد الأوروبي إلى استخدام النظم والأطر التنظيمية الحالية، بدلاً من بناء أطر تنظيمية جديدة على شركات التكنولوجيا الالتزام بها. وبالتوازي، سوف يحتاج غوغل وفيسبوك وغيرهما من المنصات التكنولوجية إلى الاستعداد لقانون الخدمات الرقمية (Digital Services Act)، الذي نتوقع صدوره بنهاية هذا العام، ويسعى إلى "تنظيم النظام الإيكولوجي على الإنترنت عبر مجموعة من المجالات منها

58 المرجع نفسه.
59 ماري شروتر (2020)، "الذكاء الاصطناعي ومكافحة التطرف العنيف"، الشبكة العالمية للتطرف والتكنولوجيا، ص.23.
60 المفوضية الأوروبية، (19 فبراير 2020)، 'White Paper on Artificial Intelligence – A European Approach to Excellence and Trust', p.2. متاح من: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
61 المرجع نفسه، ص.10.
62 المرجع نفسه، ص.20.
63 Government of France, Ministry of Europe and Foreign Affairs, 'Transparency and accountability: The challenges of artificial intelligence'. متاح من: https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/transparency-and-accountability-the-challenges-of-artificial-intelligence/
64 المرجع نفسه.
65 المرجع نفسه.

... المحتوى المسيء".[66] من المتوقع أيضًا أن يُتبِع الاتحاد الأوروبي كتابه الأبيض للذكاء الاصطناعي بتشريعات تتعلق بالذكاء الاصطناعي والسلامة والمسؤولية والحقوق الأساسية والبيانات في وقت لاحق من عام 2020.[67]

فرنسا

في فرنسا عددٌ من أصحاب المصلحة الرئيسيين الذين يقع الذكاء الاصطناعي في دائرة اختصاصاتهم. وكُلِّف المنسق الوزاري للذكاء الاصطناعي بتحليل ووضع مقترحات للتغييرات المتعلقة بالابتكار الرقمي المطبق على المجال الأمني.[68] توجد في وزارة الدفاع وحدة لتنسيق الذكاء الاصطناعي الدفاعي ضمن وكالة الابتكار الدفاعي.

تعمل فرنسا على تكييف إطارها القانوني للسماح باستخدام آمن وفعال للتقنيات التي تدعم الذكاء الاصطناعي لحماية سكان فرنسا. وفيما يتعلق بالتطورات السياسية، نشرت فرنسا استراتيجيتها للذكاء الاصطناعي في مارس 2018. ومن أهدافها الرئيسية: تحسين النظام الإيكولوجي لتعليم الذكاء الاصطناعي والتدرب عليه لتطوير وجذب أفضل مواهب الذكاء الاصطناعي؛ وإنشاء سياسة البيانات المفتوحة لتنفيذ تطبيقات الذكاء الاصطناعي وتجميع الأصول معًا؛ ووضع إطار أخلاقي لاستخدام تطبيقات الذكاء الاصطناعي بشفافية وعدل.[69] وفقًا لتوجيهات الاتحاد الأوروبي بشأن أمن الشبكات وأنظمة المعلومات، وضعت فرنسا قانون أمنها السيبراني.[70] تقوم فرنسا حاليًا بترتيب أفكارها بشأن استخدام التقنيات التي تدعم الذكاء الاصطناعي في المجال العسكري.[71]

أطلقت فرنسا سلسلة من المبادرات التي تركز على الذكاء الاصطناعي. وفي مؤتمر مجموعة السبع (G7) لأصحاب المصلحة المتعددين حول الذكاء الاصطناعي في عام 2018، أعلنت فرنسا وكندا عن إطلاق الهيئة الدولية للذكاء الاصطناعي (International Panel on Artificial Intelligence)، ما يدعم تبني الذكاء الاصطناعي بطريقة مسؤولة.[72] واشتركتا معًا في إقامة الشراكة العالمية الجديدة للذكاء الاصطناعي (GPAI) التي انضمت إليها بلدان أخرى. والهدف من هذه المبادرة توجيه تطويرات الذكاء الاصطناعي واستخداماته بطريقة مسؤولة، مع مراعاة حقوق الإنسان والاندماج والتنوع والابتكار والنمو الاقتصادي.[73] وسوف تعمل تحديدًا على سد الفجوة بين النظرية والتطبيق في الذكاء الاصطناعي بدعم البحث العلمي في أنشطة الذكاء الاصطناعي. يقوم بدعم الشراكة العالمية للذكاء الاصطناعي (GPAI) مركز التميز في فرنسا، وهو مؤسسة شقيقة لمركز تميز GPAI في مونتريال. وتدعم منظمة التعاون الاقتصادي والتنمية (OECD) الشراكة العالمية للذكاء الاصطناعي (GPAI) أيضًا.

استنادًا إلى استراتيجية الذكاء الاصطناعي، قد ترى فرنسا إصدار قانون الجمهورية الرقمية، وهذا القانون قد يعمل على "فتح البيانات العامة، وتعزيز حماية حقوق المستخدمين وخصوصية البيانات، وضمان استفادة الجميع من الفرص التي تقدمها الرقمنة".[74]

---

66 Stolton, S. (23 يونيو 2020)، 'EURACTIF Platform clamp down on hate speech in run up to Digital Services Act'. متاح من: https://www.euractiv.com/section/digital/news/platforms-clamp-down-on-hate-speech-in-run-up-to-digital-services-act/

67 Kayali, L., Heikkila, M. and Delcker, J. (19 فبراير 2020)، 'Europe's digital vision, explained', Politico. متاح: https://www.politico.eu/article/europes-digital-vision-explained/

68 حكومة فرنسا، مكتب رئيس الوزراء، (13 يوليو 2018)، 'Action plan against terrorism'، ص.20. متاح: http://www.sgdsn.gouv.fr/uploads/2018/10/20181004-plan-d-action-contre-le-terrorisme-anglais.pdf

69 المفوضية الأوروبية، تقرير استراتيجية فرنسا للذكاء الاصطناعي. متاح: https://ec.europa.eu/knowledge4policy/ai-watch/france-ai-strategy-report_en

70 Government of France, National Cyber Security Agency, Directive network and information system security (NIS). متاح: https://www.ssi.gouv.fr/entreprise/reglementation/directive-nis/

71 Pannier, A. and Schmitt, O. (2019), 'To fight another day: France between the fight against terrorism and future warfare', International Affairs vol. 95, no. 4. متاح: https://academic.oup.com/ia/article/95/4/897/5492774

72 حكومة كندا، مكتب رئيس الوزراء (6 ديسمبر 2018)، Mandate for the International Panel of Artificial Intelligence. متاح: https://pm.gc.ca/en/news/backgrounders/2018/12/06/mandate-international-panel-artificial-intelligence

73 الحكومة الفرنسية، وزارة أوروبا والشؤون الخارجية (15 يونيو 2020)، قيام 15 عضوًا مؤسِّسًا بإطلاق الشراكة العالمية للذكاء الاصطناعي. متاح: https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding

74 المفوضية الأوروبية، تقرير استراتيجية فرنسا للذكاء الاصطناعي.

غانا

لاتزال الجهود التي تبذلها غانا لمكافحة التطرف العنيف على الإنترنت محدودة، لأن العنف السياسي في غانا لم تؤججه أنشطة إرهابية، بخلاف الوضع في دولتيْ نيجيريا وتشاد المجاورتيْن لها.[75] تورد قاعدة بيانات الإرهاب العالمي، وهي قاعدة بيانات للهجمات الإرهابية العالمية منذ عام 1970، 21 حادثة فقط أسفرت عن 23 حالة وفاة خلال 50 عامًا في غانا.[76]

لا تواجه غانا نفس المشكلات التي تواجهها بعض البلدان المجاورة لها فيما يتعلق بإغلاق الحكومة للإنترنت أو الاستخدام الحكومي لوسائل التواصل الاجتماعي لقمع المعارضة السياسية.[77] استغلت هذه الحكومات إرثًا من القوانين الاستعمارية التي انتُهكت بها الحريات في الماضي، بهدف "إضفاء الشرعية على العديد من ... محاولات تقديم مطالب غير قانونية للقطاع الخاص".[78] يوضح تقرير 'Ranking Digital Rights' (تصنيف الحقوق الرقمية) لعام 2019 أن منصات وسائل التواصل الاجتماعي ومزودي خدمات الإنترنت اضطروا للاستجابة لمطالب إغلاق حكومية غير قانونية، ما أثار مخاوف من فرض المراقبة والرقابة المفرطة.[79]

مع أن الحكومة الغانية لم تقدم هذه المطالب غير القانونية حتى الآن، أعربت جماعات المجتمع المدني والصحفيون عن قلقهم بشأن المستقبل.[80] قبل انتخابات عام 2016، أعلن قائد الشرطة الغانية عن احتمال إغلاق وسائل التواصل الاجتماعي.[81] بالرغم من معارضة الرئيس هذه الخطط، تتصاعد المخاوف بشأن الحقوق الرقمية في غانا.

تترك قوانين حرية التعبير الليبرالية في غانا الفضاء الرقمي عرضةً للانتهاكات، مثل خطاب الكراهية والتنمر السيبراني (على النساء خصوصًا).[82] لذا تتزايد الدعوات لإحكام القواعد التنظيمية على منصات وسائل التواصل الاجتماعي. وأشار خبير من مؤسسة Freedom of Expression Media Foundation for West Africa (مؤسسة حرية التعبير الإعلامية لغرب أفريقيا) إلى أن "غياب القواعد التنظيمية يفضي إلى تطبيق تشريعات أخرى لمقاضاة الأفراد بطرق قد تكون مفرطة"، على غرار المطالب الحكومية الموضحة أعلاه.

ومع ذلك، يجب أن توازن اللوائح الحكومية الخاصة بوسائل التواصل الاجتماعي بين حماية المستخدمين من الأذى وحماية حرية التعبير المكفولة للمستخدمين. وحذرت إحدى جماعات المجتمع المدني، المعروفة بمناهضتها لقطع الإنترنت، من التنظيم الحكومي لوسائل التواصل الاجتماعي: "إذا تركنا للحكومة تنظيم الإنترنت – على غرار ما حدث في بلدان أخرى – انتهى بنا الأمر إلى أن تُملي علينا شروطها لاستخدام الإنترنت."[83]

لم يتضح بعد إذا كانت هناك خطط لتطوير أدوات قائمة على الذكاء الاصطناعي للمساعدة في تنظيم المحتوى عبر الإنترنت في غانا. ومع ذلك، أظهرت التهديدات الإقليمية لحرية التعبير عبر قوانين الحقبة الاستعمارية أن الدولة يجب أن تضع حقوق مواطنيها الرقمية في مقدمة أي أدوات تكنولوجية أو جهود تشريعية لمراقبة المحتوى الضار عبر الإنترنت. وفي خطوة لاقت ترحيبًا، أقرت غانا مشروع قانون الحق في المعلومات في عام 2019، والذي يضمن إمكانية الوصول إلى المعلومات التي تحتفظ

---

75 يرجع الفضل في بيان هذه الأفكار عبر البريد الإلكتروني إلى الأستاذ توميوا إيلوري، باحث في وحدة التعبير والمعلومات والحقوق الرقمية بمركز حقوق الإنسان بجامعة بريتوريا.

76 قاعدة بيانات الإرهاب العالمي، START. متاح: https://www.start.umd.edu/gtd/

77 'Content Moderation Is Particularly Hard in African Countries', Ilori, T. (2020)، مشروع لجمعية المعلومات بكلية الحقوق في ييل. متاح: https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/moderate-globally-impact-locally-content-moderation-particularly-hard-african-countries

78 'Stemming digital colonialism through reform of cybercrime laws in Africa', Ilori, T. (2020)، مشروع لجمعية المعلومات بكلية الحقوق في ييل. متاح: https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/stemming-digital-colonialism-through-reform-cybercrime-laws-africa

79 تصنيف الحقوق الرقمية، 'RDR Corporate Accountability Index 2019'. متاح: https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf

80 'Africa in urgent need of a homegrown online rights strategy', Majama, K. (2019)، جمعية الاتصالات التقدمية. متاح: https://www.apc.org/en/news/africa-urgent-need-homegrown-online-rights-strategy

81 Olukotun, D. (16 أغسطس 2019)، 'President of Ghana says no to internet shutdowns during coming elections,' AccessNow. متاح: https://www.accessnow.org/president-ghana-says-no-internet-shutdown-elections-social-media/

82 Endert, J. (2018), 'Digital backlash threatens media freedom in Ghana', DW Akademie. متاح: https://www.dw.com/en/digital-backlash-threatens-media-freedom-in-ghana/a-46602904

83 المرجع نفسه.

بها المؤسسات العامة.[84] يشير مشروع القانون إلى أن الحكومة الغانية تريد التعامل مع الحقوق الرقمية بشفافية ومساءلة. وأي تطويرات مستقبلية في مجال إدارة المحتوى عبر الإنترنت ينبغي أن تستوفي هذه الالتزامات والمعايير لحماية الخصوصية وحقوق حرية التعبير.

اليابان

توجه اليابان معظم جهود مكافحتها التطرف العنيف من خلال رابطة دول جنوب شرق آسيا (آسيان).[85] في مطلع عام 2004، أصدرت دول الآسيان الأعضاء، بالشراكة مع اليابان، مجموعة من البيانات التوضيحية للتعاون في مكافحة الإرهاب الدولي. وفضلاً عن الإشارة إلى النوايا السياسية، ألزم الإعلان الموقعين عليه "بمنع الإرهاب الدولي وتعطيله ومكافحته بتبادل المعلومات وتبادل الاستخبارات وبناء القدرات"، ما يُعد سابقة في التعاون الإقليمي لمكافحة التطرف العنيف والإرهاب.[86]

وأكدت اليابان مجددًا التزامها بالتعاون متعدد الجنسيات في جنوب شرق آسيا في عام 2015 في مكافحة الإرهاب والتطرف العنيف والتعاون في تنفيذ خطة عمل الآسيان لمنع ومكافحة تصاعد وتيرة الراديكالية والتطرف العنيف (2018–2025).[87] تعطي خطة العمل أولوية للشراكة "مع مجتمع الأعمال وقطاع التكنولوجيا في تعزيز الاعتدال وتدعيم الحوار لمنع الراديكالية والتطرف العنيف"، فضلاً عن تعزيز "الاتصالات الاستراتيجية" لمنع الإرهابيين والمتطرفين العنيفين من إساءة استخدام وسائل التواصل الاجتماعي.[88]

سوف تقام دورة الألعاب الأولمبية والبارالمبية لعام 2020 (التي تأجلت حتى عام 2021 بسبب أزمة فيروس كورونا) في طوكيو. جرت العادة أن يُنظر إلى استضافة الألعاب الأولمبية على أنها "اختبار" للقدرات الأمنية للدولة المعنية، وتعد الألعاب فرصة لتجريب الابتكارات في مجال الذكاء الاصطناعي والأمن وإنفاذ القانون.[89]

من هذه التجارب قبل الألعاب في عام 2018، أعلنت شرطة محافظة كاناغاوا عن إطلاق نظام الشرطة التنبؤية للتنبؤ بالجرائم والهجمات على أساس خوارزمية التعلم الآلي العميقة.[90] منذ ذلك الحين، أكدت شركات التكنولوجيا الرائدة توفير إمكانية التعرف على الوجوه على نطاق واسع، والمصادقة البيومترية، وأنظمة كشف السلوك في الألعاب وفي الموانئ والمطارات.[91] تمتاز هذه الأنظمة بقدرتها على فحص الوجوه بحثًا عن مشاعر معينة وتأكيد هويات أصحابها على أساس ميزات الوجه والمعلومات الشخصية.

لم يتضح بعد إذا كانت قدرات الذكاء الاصطناعي الأمنية هذه سوف تمتد إلى وسائل التواصل الاجتماعي والنشاط عبر الإنترنت. وبحسب ما ورد، كان من الممكن أن تتضمن الأنظمة التي تمت تجربتها في كاناغاوا مراقبة محتوى وسائل التواصل الاجتماعي لمكافحة الجريمة، الأمر الذي يمكن فهمه كترحيب من سلطات إنفاذ القانون اليابانية بمراقبة وسائل التواصل الاجتماعي القائمة على الذكاء الاصطناعي لمكافحة المحتوى

---

84 Yahya Jafru, M. (26 مارس 2019)، 'Right to information – RTI bill passed into law', Graphic Online'. متاح: https://www.graphic.com.gh/news/politics/ghana-news-rti-bill-passed.html

85 'Japan: Extremism & Counter Extremism', Counter-Extremism Project. متاح: https://www.counterextremism.com/countries/japan

86 'ASEAN-Japan Joint Declaration for Cooperation to Combat International Terrorism', ASEAN. متاح: https://asean.org/?static_post=asean-japan-joint-declaration-for-cooperation-to-combat-international-terrorism-2

87 'Chairman's Statement of the 18th ASEAN-Japan Summit, Kuala Lumpur, November 22 2015', ASEAN. متاح: https://www.asean.org/wp-content/uploads/2015/12/6.-Chairmans-Statement-of-the-18th-ASEAN-Japan-Summit-Final-Final.pdf;
'Japan's cooperation with ASEAN 2025 (Counter-Terrorism)', Mission of Japan to ASEAN. متاح: https://www.asean.emb-japan.go.jp/asean2025/jpasean-ps03.html

88 '2018 ASEAN Plan of Action to Prevent and Counter the Rise of Radicalisation and Violent Extremism (2018–2025), adopted in Myanmar, October 31 2018', ASEAN. متاح: https://cil.nus.edu.sg/wp-content/uploads/2019/10/2018%20ASEAN%20Plan%20of%20Action%20to%20Prevent%20and%20Counter%20the%20Rise%20of%20Radicalisation%20and%20Violent%20Extremism%20(2018-2025).pdf

89 انظر، على سبيل المثال: Soria, V. (2011), 'Beyond London 2012: The Quest for a Security Legacy,' The RUSI Journal, vol. 156, no. 2, pp.36–43.

90 'Kanagawa police to launch AI-based predictive policy system before Olympics', Japan Times (29 يناير 2018). متاح [غير مجاني]: https://www.japantimes.co.jp/news/2018/01/29/national/crime-legal/kanagawa-police-launch-ai-based-predictive-policing-system-olympics/

91 'All is Ready for a Safe and Secure Tokyo Games' The Government of Japan (2019). متاح: https://www.japan.go.jp/tomodachi/2019/autumn-winter2019/tokyo2020.html;
'NEC Becomes a Gold Partner for the Tokyo 2020 Olympic and Paralympic Games', NEC Corporation (2015). متاح: https://www.nec.com/en/press/201502/global_20150219_01.html

الضار أو الذي يحتمل أن يكون خطيرًا على الإنترنت. وقد تكون هذه الخطوة محفوفة بالمخاطر، بالنظر إلى احتجاجات عام 2017 ردًا على مشروع قانون الحكومة المثير للجدل لمكافحة الإرهاب، والذي رأى النقاد أنه يخاطر بالحريات المدنية.[92] سوف تحتاج اليابان إلى إيلاء اهتمام جاد لحماية حقوق المواطنين في الخصوصية وحرية التعبير بينما تقوم بتطوير تقنيات الذكاء الاصطناعي لمكافحة التطرف العنيف عبر الإنترنت وعلى أرض الواقع.

نيوزيلندا

تتضمن حوكمة مكافحة التطرف العنيف عبر الإنترنت في نيوزيلندا التنسيق بين مختلف الوكالات والهيئات. ومنها لجنة العلاقات الخارجية والأمن التابعة لمجلس الوزراء؛ وكالات الشرطة والاستخبارات والاتصالات الأمنية؛ ووكالات الشؤون الخارجية والتجارة والدفاع والنقل والابتكار والتنمية. وأدرجت استراتيجية نيوزيلندا الشاملة في تعميم استراتيجيتها الوطنية، الصادرة في فبراير 2020.[93]

في أعقاب إطلاق النار على المصلين في مسجد كرايستشيرش في مارس 2019، كوَّنت حكومتا نيوزيلندا وفرنسا تحالفًا من رؤساء الدول مع شركات وسائل التواصل الاجتماعي والتكنولوجيا في إطار دعوة كرايستشيرش للقضاء على المحتوى الإرهابي والعنف والمحتوى المتطرف عبر الإنترنت.[94] وتلزم هذه الدعوة الدول الداعمة لها بإنفاذ القوانين التي تحظر نشر المحتوى الإرهابي والمتطرف العنيف عبر الإنترنت مع احترام القانون الدولي لحقوق الإنسان بما في ذلك حرية التعبير. وتعمل هذه البلدان أيضًا على دعم الأطر وأنشطة بناء القدرات وتعزيز الوعي لمنع استغلال الخدمات عبر الإنترنت لنشر المحتوى الإرهابي والمتطرف العنيف.

كما تلزم دعوة كرايستشيرش الشركات، ومنها أمازون وفيسبوك وغوغل وتويتر ويوتيوب، بمعايير الصناعة الأكبر التي تعزز المساءلة والشفافية. ويجب على الشركات فرض معايير المجتمع وشروط الخدمة بإيلاء الأولوية لعمليات إدارة المحتوى وإزالته، والتعرف على المحتوى لحظة بلحظة لمراجعته وتقييمه. وتعمل البلدان والشركات معًا على تطوير جهودها مع المجتمع المدني لتعزيز أنشطة المجتمع المدني للتدخل في عمليات الراديكالية عبر الإنترنت.

وبعد هجوم مارس 2019، تم إنشاء مفوضية تحقيق ملكية (Royal Commission of Inquiry) لتقييم استجابة الوكالات لحادث إطلاق النار وتحديد التدابير الأخرى التي يمكن اتخاذها لمنع الهجمات في المستقبل.[95] سوف يسلط تقرير المفوضية الضوء على استراتيجية مكافحة التطرف العنيف الحالية وتوجهها المستقبلي في نيوزيلندا، وسوف تكون موردًا مفيدًا لفهم نصيب الذكاء الاصطناعي في هذه الاستراتيجية المستقبلية. وتأخر نشر التقرير حتى شتاء 2020، بسبب أزمة فيروس كورونا.

تلتزم حكومة نيوزيلندا أيضًا بمعايير أشد صرامة للشفافية والمساءلة في استخدام خوارزميات الحوكمة. وكما ورد في الذكاء الاصطناعي ومكافحة التطرف العنيف: كتاب تمهيدي فإن الاستخدام الخوارزمي قد يفاقم التحيزات الموجودة فعلاً.[96] في يوليو 2020، نشرت الحكومة ميثاق الخوارزمية لأوتياروا نيوزيلندا (Algorithm Charter for Aotearoa New Zealand)، ويُعد مراجعةً شاملةً لاستخدام الدولة للخوارزميات في قطاعات تتراوح من النقل إلى العدالة، والتزامًا بمزيد من الشفافية ومشاركة أصحاب

92 'Japan passes controversial anti-terror conspiracy law', BBC (15 يونيو 2017). متاح: https://www.bbc.co.uk/news/world-asia-40283730

93 حكومة نيوزيلندا، لجنة المسؤولين لتنسيق الأمن الداخلي والخارجي، لجنة تنسيق مكافحة الإرهاب (فبراير 2020)، 'Countering terrorism and violent extremism national strategy overview'. https://dpmc.govt.nz/sites/default/files/2020-02/2019-20 CT Strategy-all-final.pdf

94 انظر https://www.christchurchcall.com/

95 المفوضية الملكية للتحقيق في الهجوم على مسجدَيْ كرايستشيرش، انظر: https://christchurchattack.royalcommission.nz/

96 انظر أيضًا RUSI, 'Briefing Paper: Data Analytics and Algorithmic Bias in Policing', Babuta, A. and Oswald, M. (2019). متاح: https://www.gov.uk/government/publications/report-commissioned-by-cdei-calls-for-measures-to-address-bias-in-police-use-of-data-analytics;
Benjamin, R. (2019), Race After Technology: Abolitionist Tools for the New Jim Code (Polity);
Benjamin, R., 'A New Jim Code?', Berkman Klein Center for Internet & Society at Harvard University. التسجيل متاح: https://cyber.harvard.edu/events/new-jim-code

المصلحة، وضمانات للخصوصية والإشراف البشري على استخدام الخوارزميات.[97] بلغ عدد الموقعين على الميثاق – وهو الأول من نوعه على مستوى العالم – خمسًا وعشرين هيئة حكومية وقت كتابة هذا التقرير.

لكن الوكالات والهيئات المسؤولة عن مكافحة التطرف العنيف على الإنترنت غير موجودة فعليًا. ولتحقيق هذا الهدف، لم يتضح بعد إلى أي مدى يفكر صانعو السياسات في نيوزيلندا في تطوير أدوات الذكاء الاصطناعي والخوارزميات لمواجهة المحتوى الضار على الإنترنت والمعايير التي تلتزم بها هذه الأدوات. ويُعد الميثاق خطوة في الاتجاه الإيجابي، ويعتبر تطبيق هذه المعايير على إجراءات مكافحة التطرف العنيف القائمة على الذكاء الاصطناعي تطورًا مرحبًا به في صنع السياسات.

المملكة المتحدة

في فبراير 2018، أعلنت المملكة المتحدة عن تطوير أداة خوارزمية قائمة على التعلم الآلي لكشف محتوى داعش الإرهابي عبر الإنترنت. وتم تدريب البرنامج على تحديد العناصر السمعية والبصرية التي يمكن التعرف عليها في المحتوى الدعائي لتنظيم الدولة والإشارة إليها – الرايات والشعارات والتنسيق والهياكل والموسيقى التصويرية. واستثمرت منصات التكنولوجيا العملاقة مثل يوتيوب وفيسبوك استثمارًا كبيرًا في تطوير أدوات إدارتها الآلية للمحتوى على مر السنين. وصُممت الأداة بطريقة مستقلة عن أي منصة معينة، وبالتالي فهي مفتوحة المصدر ويمكن استخدامها في منصات الإنترنت والتواصل الاجتماعي الأصغر مثل Vimeo.

ومع أنها أداة واعدة، نجد فعاليتها وقبولها محدود للغاية. أولاً، كما أوضح زميلنا تشارلي وينتر، يتراوح محتوى داعش عبر الإنترنت من مقاطع الفيديو إلى الصور الفوتوغرافية والمقاطع المكتوبة والنشرات الإذاعية. وتُعد مكافحة محتوى الفيديوهات خطوة إيجابية، "تخفف [المشكلة] نوعًا ما، في أحسن الأحوال، لكنها أبعد ما يمكن عن الحل."[98] ثانيًا، قامت وزارة الداخلية البريطانية بتكليف الأداة بالتعرف على أشد فيديوهات داعش جرأةً وترويعًا. وأوضحت شركة تطوير البرمجيات التي صممت الأداة أن "الأمر لا يتعلق بالحجم بقدر ما يتعلق بمدى تأثيرها حسب اعتقادهم [وزارة الداخلية البريطانية] في معالجة مجموعات معينة من مقاطع الفيديو".[99] ومع ذلك، وصفت العديد من الدراسات الأكاديمية التأثير الواسع النطاق للمحتوى الدعائي "الألين" وقدراته الراديكالية على مدى فترات زمنية طويلة.[100] تقييد الذكاء الاصطناعي وتضييق نطاق وظائفه بثلاث طرق – تنظيم داعش ومحتوى الفيديو والمحتوى المتطرف – يضعف الفعالية التقنية للأداة. [101] مع أن الأداة أتيحت مجانًا لشركات التكنولوجيا الأصغر، لم تعتمدها، اعتبارًا من أبريل 2019، أي شركة بعد.

يوضح نهج حكومة المملكة المتحدة في استخدام الذكاء الاصطناعي لمكافحة التطرف العنيف عبر الإنترنت أيضًا احتمالية ظهور تضارب في المصالح أثناء التعاون بين الحكومات وهذه الصناعة. حدد كتاب الحكومة الأبيض بشأن الأضرار على الإنترنت (Online Harms White Paper)، الذي نُشر في أبريل 2019، حالة نموذجية شاملة لمزيد من التنظيم الوطني لوسائل التواصل الاجتماعي.[102] وحسب هذا الإطار التنظيمي الجديد، سوف تتحمل وسائل التواصل الاجتماعي وشركات التكنولوجيا واجبًا قانونيًا جديدًا يتمثل في رعاية مستخدميها، وتتولى Ofcom تنفيذه بصفتها الهيئة التنظيمية للاتصالات في المملكة المتحدة. وسوف تفرض هيئة Ofcom على المنصات عقوبات مالية وتقنية – يمكن حظر مواقع الويب على مستوى مزود خدمة الإنترنت وفرض غرامة تصل إلى 4٪ من إيراداتها العالمية – عند عدم الامتثال لإطار العمل وانتهاك

97 'Algorithm charter for Aotearoa New Zealand', data.govt.nz. متاح: https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter

98 Temperton, J. (13 فبراير 2018)، 'ISIS could easily dodge the UK's AI-powered propaganda blockade', Wired. متاح: https://www.wired.co.uk/article/isis-propaganda-home-office-algorithm-asi

99 المرجع نفسه.

100 'Hashtag Terror: How ISIS Manipulates Social Media', Anti-Defamation League (21 أغسطس 2014). متاح: https://www.adl.org/education/resources/reports/isis-islamic-state-social-media

101 Murgia, M. and Bond, D. (6 أبريل 2019)، 'Businesses show no appetite for anti-terror AI tool', Financial Times. متاح [غير مجاني]: https://www.ft.com/content/fda2d218-56fb-11e9-91f9-b6515a54c5b1

102 حكومة جلالتها (أبريل 2019)، 'Online Harms White Paper'. متاح: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

واجب الرعاية القانوني.[103] وعند الإعلان عن الأداة الخوارزمية في فبراير 2018، أشار وزير الداخلية آنذاك أمبر رود إلى أن القانون قد يلزم الشركات بتبني هذه الأداة.

وهذه التحركات التنظيمية ليست مقلقة في حد ذاتها. ومع ذلك، كُلفت شركة تحليل البيانات وتطوير البرمجيات التي طورت هذه الأداة، المعروفة سابقًا باسم ASI Data Science (تعرف الآن باسم Faculty)، بنمذجة البيانات في حملات Vote Leave and Leave.EU، وهكذا تورطت في فضيحة Cambridge Analytica.[104] علاوة على ذلك، اعتبارًا من مايو 2020، مُنحت الشركة ما لا يقل عن سبعة عقود حكومية يمولها القطاع العام في ثمانية عشر شهرًا، ولديها علاقات شخصية وتجارية جديرة بالملاحظة مع دومينيك كامينغز، كبير مستشاري رئيس الوزراء.[105]

وتثير هذه الحقائق مخاوف من تضارب المصالح. وتهتز ثقة الجمهور ورواد الأعمال في الأداة بمنح عقد تطوير الأداة إلى شركة لها صلات وثيقة بالدوائر الداخلية للحكومة، وكانت متورطة في فضيحة عامة، ناهيك عن دعم التشريعات التي قد تلزم منصات وسائل التواصل الاجتماعي باستخدامها لضمان استمرار أعمال الشركة التجارية.

وكان من الممكن أن تقوم أي شركة تطوير مستقلة عن الحكومة بتطوير أداة تقنية أكثر فعالية تلبي متطلبات أوسع نطاقًا (تستطيع مثلاً التعرف على أكثر من مجموعة فرعية من محتوى فيديوهات تنظيم داعش، وتعزيز الثقة في الأداة والإقبال عليها. والشفافية والمساءلة "ليستا مجرد شعاريْن لخطب الود: ولكنهما بالغا الأهمية لنجاح جهود حل المشكلات" عند مكافحة التطرف العنيف عبر الإنترنت باستخدام الذكاء الاصطناعي.[106] لقد أضاعت المملكة المتحدة فرصةً سياسيةً كبيرةً لتطوير وتوفير أحدث تقنيات الذكاء الاصطناعي في مجال إدارة المحتوى الضار عبر الإنترنت عندما قوضت الثقة في الأداة وضحَّت بفعاليتها التقنية.

المديرية التنفيذية للجنة الأمم المتحدة لمكافحة الإرهاب

أنشِئت المديرية التنفيذية للجنة مكافحة الإرهاب التابعة للأمم المتحدة (UN CTED) بموجب قرار مجلس أمن الأمم المتحدة رقم 1535 (2004) كهيئة خبراء لدعم لجنة مكافحة الإرهاب التابعة لمجلس الأمن.[107] وكان هدفها الأولي تقييم تنفيذ الدول الأعضاء في الأمم المتحدة لقرارات مجلس الأمن بشأن مكافحة الإرهاب ودعم جهودها بالحوار. وتعمل المديرية التنفيذية للجنة مكافحة الإرهاب التابعة للأمم المتحدة (UN CTED) عن كثب مع مجلس الأمن وشركات التكنولوجيا الكبرى ومنظمات المجتمع المدني من خلال منتدى الإنترنت العالمي لمكافحة الإرهاب (GIFCT).

يوجد حاليًا العديد من قرارات مجلس الأمم المتحدة بشأن إساءة استخدام الإنترنت لأغراض إرهابية، وتسعى المديرية التنفيذية للجنة مكافحة الإرهاب التابعة للأمم المتحدة (UN CTED) إلى تعزيز الاتساق وتيسير سبل التكامل بين قرارات مجلس أمن الأمم المتحدة ودور تكنولوجيا المعلومات. ويعترف قرار مجلس الأمن 2129 (2013) بتزايد العلاقة بين الإرهاب وتكنولوجيا المعلومات، واستخدام تقنيات حديثة مثل الإنترنت في ارتكاب الأعمال الإرهابية وتسهيلها، من خلال السماح بالتحريض على الأعمال الإرهابية

103 Crawford, A. (29 يونيو 2020)، 'Online Harms bill: Warning over "unacceptable" delay', BBC. متاح: https://www.bbc.co.uk/news/technology-53222665

104 Evans, R. and Pegg, D. (4 مايو 2020)، 'Vote Leave AI firm wins seven government contracts in 18 months', The Guardian. متاح: https://www.theguardian.com/world/2020/may/04/vote-leave-ai-firm-wins-seven-government-contracts-in-18-months;
Pegg, D., Evans, R. and Lewis, P. (12 يوليو 2020)، 'Revealed: Dominic Cummings firm paid Vote Leave's AI firm £260,000', The Guardian. متاح: https://www.theguardian.com/politics/2020/jul/12/revealed-dominic-cummings-firm-paid-vote-leaves-ai-firm-260000;
Pegg, D. and Evans, R. (2 يونيو 2020)، 'AI firm that worked with Vote Leave given new coronavirus contract' The Guardian. متاح: https://www.theguardian.com/technology/2020/jun/02/ai-firm-that-worked-with-vote-leave-wins-new-coronavirus-contract

105 Cadwalladr, C. (7 مايو 2017)، 'The great British Brexit robbery: how our democracy was hijacked', The Guardian. متاح: https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy

106 'Tackling the Information Crisis: A Policy Framework for Media System Resilience', The Report of the LSE Commission on Truth Trust & Democracy, p.32. متاح: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

107 Chowdhury Fink, N. (2012), 'Meeting the challenge: A guide to United Nations counterterrorism activities', International Peace Institute, p. 45. https://www.ipinst.org/wp-content/uploads/publications/ebook_guide_to_un_counterterrorism.pdf

أو التجنيد أو جمع الأموال أو التخطيط لها.[108] ويعيد هذا القرار أيضًا إنفاذ ولاية المديرية التنفيذية للجنة مكافحة الإرهاب التابعة للأمم المتحدة (UN CTED). وتطالب القرارات 2354 (2017) و 2395 (2017) و 2396 (2017) الدول الأعضاء بالأمم المتحدة بالتعاون لمنع المنظمات الإرهابية من استغلال تكنولوجيا المعلومات والاتصالات، والعمل مع القطاع الخاص والمجتمع المدني لوضع تدابير فعالة لمنع إساءة استخدام الإنترنت لأغراض إرهابية.[109] يدعو قرار مجلس الأمن رقم 1373 الدول الأعضاء بالأمم المتحدة إلى تطوير وتسريع "تبادل المعلومات العملياتية" حول استخدام المنظمات الإرهابية لتكنولوجيا المعلومات ووقف تجنيد الإرهابيين.[110]

تسعى الفريق الرفيع المستوى التابع للأمين العام للأمم المتحدة المعني بالتعاون الرقمي إلى إيجاد حلول لتخفيف مخاطر الذكاء الاصطناعي.[111] تنص توصية الفريق 3ج على الآتي:

> نعتقد أن الأنظمة الذكية المستقلة ينبغي تصميمها بطرق تمكِّن من شرح قراراتها وأن يكون البشر مسؤولين عن استخدامها. وينبغي أن تراقب عمليات التدقيق وخطط الاعتماد امتثال أنظمة الذكاء الاصطناعي للمعايير الهندسية والأخلاقية، والتي ينبغي تطويرها وفقًا لمناهج أصحاب المصلحة المتعددين ومتعددة الأطراف. وينبغي عدم تفويض الآلات باتخاذ قرارات تعني الحياة أو الموت. وندعو إلى تعزيز التعاون الرقمي مع العديد من أصحاب المصلحة للتفكير في تصميم وتطبيق هذه المعايير والمبادئ مثل الشفافية وعدم التحيز في الأنظمة الذكية المستقلة في بيئات اجتماعية مختلفة.[112]

من بين مجالات التركيز حماية حقوق الإنسان في العصر الرقمي.[113]

الولايات المتحدة

تحدد إستراتيجية الولايات المتحدة لمكافحة الإرهاب مكافحة التطرف العنيف على الإنترنت مجالاً ذا أولوية، وتلتزم بالعمل مع قطاع الأعمال والصناعة لمكافحة عمليات تجنيد الإرهابيين وجمع الأموال والراديكالية عبر الإنترنت. فيما يتعلق بالمبادرات عبر الوطنية، تعمل الولايات المتحدة الأمريكية مع مبادرات مثل Tech Against Terrorism (التكنولوجيا في مواجهة الإرهاب) والمنتدى العالمي لمكافحة الإرهاب، الذي يعتمد على الشراكة مع الموقعين الآخرين والمجتمع المدني وقطاع التكنولوجيا لصياغة نهج لمكافحة التطرف العنيف عبر الإنترنت على المدى المتوسط والبعيد.

وفي المجال التشريعي المحلي للولايات المتحدة الأمريكية، انطلقت دعوات لتنظيم وسائل التواصل الاجتماعي ومنصات التكنولوجيا مدفوعة بتقارير عن التدخل الروسي وعمليات المعلومات عبر الإنترنت في الانتخابات الرئاسية لعام 2016. وفي الوقت ذاته، واصلت شركات وسائل التواصل الاجتماعي نموها من حيث أعداد المستخدمين والخدمات والمنتجات الفرعية. وفي أواخر عام 2019، عقدت اللجنة المصرفية بمجلس الشيوخ الأمريكي (US Senate Banking Committee) ولجنة الكونجرس للطاقة والتجارة (Congressional Committee on Energy and Commerce) جلسات استماع

108 الأمم المتحدة، لجنة مكافحة الإرهاب التابعة لمجلس الأمن، (14 سبتمبر 2018)، 'Public-private efforts to address terrorist content online: A year of progress – what's next?'. متاح: https://www.un.org/sc/ctc/news/event/public-private-efforts-address-terrorist-content-online-year-progress-whats-next/؛
المبادرة العالمية لمكافحة الجريمة المنظمة عبر الوطنية، Responding to terrorist use of the internet (21 يناير 2019). متاح: https://globalinitiative.net/terrorist_use_internet/

109 الأمم المتحدة، لجنة مكافحة الإرهاب التابعة لمجلس الأمن، 2018.

110 المبادرة العالمية لمكافحة الجريمة المنظمة عبر الوطنية، 2019.

111 Government of France, Ministry of Europe and Foreign Affairs, 'Transparency and accountability: The challenges of artificial intelligence'.

112 الأمم المتحدة (16 ديسمبر 2019)، High-level panel follow-up roundtable 3C – Artificial Intelligence – Meeting note. متاح: https://www.un.org/en/pdfs/HLP%20Followup%20Roundtable%203C%20Artificial%20Intelligence%20-%201st%20Session%20Summary.pdf

113 الأمم المتحدة، الفريق الرفيع المستوى التابع للأمين العام المعني بالتعاون الرقمي. متاح: https://www.un.org/en/digital-cooperation-panel/

بشأن خدمة العملة المشفرة ليبرا (Libra) التي اقترحها فيسبوك. وكانت جلسات الاستماع فرصة للمشرعين في الولايات المتحدة الأمريكية لاستجواب المسؤولين التنفيذيين في فيسبوك بشأن التلاعب بمنصتها وإساءة استخدامها،[114] ولإقرار التنظيم التقني الكبير كخيار تشريعي قابل للتطبيق.[115]

وفيما ينظر الكونجرس في اتخاذ إجراء تشريعي، يقوم مجتمع الاستخبارات الأمريكية بدفع استخدام الذكاء الاصطناعي قدمًا لمواجهة التطرف العنيف على الإنترنت. وفي ربيع وصيف 2019، اهتزت الولايات المتحدة الأمريكية بموجة عمليات إطلاق نار جماعي كان لمرتكبيها تاريخٌ طويلٌ مع التطرف العنيف عبر الإنترنت. مثلاً، نشر مطلق النار في كنيس بواي، الذي أطلق النار على كنيس يهودي في كاليفورنيا في أواخر أبريل 2019، بيانًا على 8chan قبل الهجوم بوقت قصير. ويشير البيان إلى عمليات إطلاق نار أخرى عبر الإنترنت، مثل إطلاق النار على المصلين في مسجد كرايستشيرش وإطلاق النار على كنيس بيتسبرغ، وإلى شخصيات ومراجع نموذجية من اليمين المتطرف والقوميين البيض على الإنترنت.

وفي هذا السياق، أعلن مكتب التحقيقات الفيدرالي (FBI) عن مناقصة لمقاولي القطاع الخاص لتطوير تكنولوجيا تمكّن المكتب "من الوصول شبه الفوري إلى مجموعة كاملة من عمليات تبادل المعلومات عبر وسائل التواصل الاجتماعي" وذلك "لكشف وتعطيل والتحقيق في مجموعة كبيرة ومتنوعة من التهديدات المتزايدة للمصالح القومية الأمريكية."[116] وطُرحت مناقصة مماثلة في يناير 2020.[117] في يونيو 2020، حينما اجتاحت احتجاجات #BlackLivesMatter جميع أنحاء البلاد، مدَّد مكتب التحقيقات الفيدرالي عقوده مع شركة Dataminr، وهي شركة لمراقبة وتحليل وسائل التواصل الاجتماعي، وشركة Venntel المعنية ببيانات المواقع.[118]

وتمضي هذه التقنيات وإمكانية الوصول إلى البيانات بالتوازي مع نظام الذكاء الاصطناعي العام كما أوضحنا في القسم الرابع المعني بالذكاء الاصطناعي ومكافحة التطرف العنيف: كتاب تمهيدي، نظام تنبؤي يتيح لتطبيق القانون إمكانية التدخل بناءً على آلية تنبيه معينة. وسوف تمثّل هذه الأدوات تهديدًا أخلاقيًا ملحوظًا لحق المستخدمين في الخصوصية، لأن مراقبة السلوك الفردي لحظة بلحظة لإنفاذ القانون تعتمد على بيانات غير مجهولة المصدر. وقد يقوِّض جمع البيانات التي يمكن التعرف عليها حق الإنسان في الأمن والأمان وحماية الهوية وحرية التعبير.

ويوضح نهج الولايات المتحدة الأمريكية في استخدام الذكاء الاصطناعي لمكافحة التطرف العنيف عبر الإنترنت التحديات القانونية والأخلاقية التي ينطوي عليها تتبع المواد عبر الإنترنت وإدارتها. وهذا النهج، كما ذكرت ماري شروتر، "لن يزيدنا إلا مرارة وألمًا."[119]

114 US House of Representatives Committee on Energy and Commerce, 'Facebook: Transparency and Use of Consumer Data,' transcript of 11 April 2018, p.33. متاح: https://docs.house.gov/meetings/IF/IF00/20180411/108090/HHRG-115-IF00-Transcript-20180411.pdf

115 Molla, R. and Stewart, E. (2019), 'How 2020 Democrats think about breaking up Big Tech', Vox. متاح: https://www.vox.com/policy-and-politics/2019/12/3/20965447/tech-2020-candidate-policies-break-up-big-tech

116 US Government Federal Acquisitions Service, 'Contract Opportunity: Social Media Alerting Subscription.' متاح: https://beta.sam.gov/opp/b6de554012cf4ab9ab795f52c638467c/view

117 US Government Federal Acquisitions Service, 'Request for Proposal – FBI Social Media Alerting.' متاح: https://beta.sam.gov/opp/2b3003e9b0b34b639687786e8420013b/view

118 US Government Federal Acquisitions Service, 'Contract Information – Dataminr. Inc.' متاح: https://beta.sam.gov/entity/962138942?keywords=Dataminr&sort=-relevance&index=&is_active=true&page=1&status=active; Fang, L. (24 يونيو 2020)، 'FBI Expands Ability to Collect Cellphone Location Data, Monitor Social Media, Recent Contracts Show', The Intercept. متاح: https://theintercept.com/2020/06/24/fbi-surveillance-social-media-cellphone-dataminr-venntel/

119 ماري شروتر (2020)، "الذكاء الاصطناعي ومكافحة التطرف العنيف"، الشبكة العالمية للتطرف والتكنولوجيا، ص.20.

توصيات السياسات

تطرح المبادرات والإجراءات الحالية، كالمبينة أعلاه، رؤى وتوصيات لصانعي السياسات في جميع أنحاء العالم. وبناءً على النتائج التي توصلنا إليها، نقدم توصيات السياسات التالية:

التوصية 1: إنشاء هيئة تنظيمية مستقلة على المستوى الدولي للإشراف على الجهود الوطنية لمكافحة التطرف العنيف عبر الإنترنت باستخدام الذكاء الاصطناعي

التشريعات الحكومية التي تفرض عقوبات على شركات التواصل الاجتماعي التي تخفق في تعديل المحتوى الضار[120] يمكنها أن تكون فعالة جدًا،[121] لكنها تخاطر بتقييد حق المواطنين في حرية التعبير، لأن الخوف من فرض العقوبات قد يؤدي إلى الإفراط في إزالة المحتوى. وكما أوضحنا أعلاه، في حالات المملكة المتحدة واليابان والولايات المتحدة الأمريكية، فإن جهود إدارة المحتوى وتشريعاتها يمكنها أن تصطدم بقضايا قانونية وأخلاقية تتعلق بالخصوصية والثقة والمساءلة.

قد تظهر فعالية التنظيم الذاتي لوسائل التواصل الاجتماعي، حيث تقوم الشركات بوضع وتطبيق معاييرها وقواعدها وسياساتها الخاصة لإزالة المحتوى الضار عبر الإنترنت، ولكن ربما يفتقر تطبيق المعايير إلى الاتساق والشفافية.[122] تنشر العديد من الشركات الكبرى بيانات عالية المستوى عن إدارة المحتوى، وليس واجبًا عليها القيام بذلك.[123]

ينبغي أن تكون عملية التنظيم مشتركة بين الحكومة والمجتمع المدني والصناعة، تحت إشراف هيئة دولية مستقلة لضمان الالتزام بالمعايير في المساءلة والشفافية والأخلاقيات. ووجود هيئة مستقلة ملتزمة بالمعايير العالمية لحماية الخصوصية،[124] ولديها آليات تنفيذية، ومستقلة عن الحكومة، يخفف من هذه المشكلات.

ومشاركة الحكومة ومنصات التواصل الاجتماعي والمجتمع المدني في عملية التنظيم تضمن وضع مصالح المستخدمين في طليعة جهود التنظيم. والتشريع الحكومي الذي يفوض الهيئة التنظيمية يحميها من تحول المصالح السياسية ويعزز آليات إنفاذها. وإلزام المنصات بالامتثال لهذه الآليات يعزز العدل والإنصاف في تطبيق جهود إدارة المحتوى. وينبغي أن تستند قيم الهيئة ومبادئ حوكمتها إلى الخصوصية وحرية التعبير والمساءلة.

التوصية 2: إدراج تدابير مكافحة التحيز الخوارزمي في عمليات تطوير البرمجيات عند التصميم

نادرًا ما تتضمن سياسات وممارسات إدارة المحتوى وتنظيمه بواسطة المنصات عبر الإنترنت مساءلة للجمهور أو مساهمات من الجمهور.[125] قد تكون الخوارزميات، التي يتم تطويرها غالبًا "خلف أبواب مغلقة" في صناعة التكنولوجيا ولها تأثير كبير على تجارب مليارات المستخدمين عبر الإنترنت، شديدة التحيز. الخوارزميات "تتعلم من خلال تغذيتها بصور معينة يختارها في الغالب مهندسون" أغلبيتهم الساحقة غير المتكافئة بيضٌ

120 على سبيل المثال، يفرض القانون الألماني لامتثال الشبكات لعام 2017، المعروف باسم NetzDG، عقوبات تصل إلى 50 مليون يورو على منصات التواصل الاجتماعي التي لا تزيل المحتوى غير القانوني في غضون أربع وعشرين ساعة. انظر: http://wp.ceps.eu/wp-content/uploads/2018/11/RR%20No2018-09_Germany's%20NetzDG.pdf

121 Elhai, W. (2020), 'Regulating Digital Harm Across Borders: Exploring a Content Platform Commission', '20SMSociety: المؤتمر الدولي لوسائل التواصل الاجتماعي والمجتمع، https://doi.org/10.1145/3400806.3400832, pp.223–4.

122 انظر Matsakis, L. (2 مارس 2018)، 'YouTube Doesn't Know Where Its Own Line Is', Wired. متاح: https://www.wired.com/story/youtube-content-moderation-inconsistent/

123 المرجع نفسه.

124 مثل مبادرة الشبكة العالمية. انظر https://globalnetworkinitiative.org/

125 'Tackling the Information Crisis: A Policy Framework for Media System Resilience,' The Report of the LSE Commission on Truth Trust & Democracy, p18. متاح: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

وذكورٌ.[126] أفضى هذا التحيز إلى مشكلات كبيرة على أرض الواقع، بما في ذلك البرامج التي تصنف المتهمين السود على أنهم أكثر عرضة لارتكاب الجرائم في المستقبل،[127] وتطبيق Google Photo الذي وسم صور مستخدم أسود بغوريلا عن طريق الخطأ.[128]

فيما يتعلق بمكافحة التطرف العنيف، هذا التحيز الخوارزمي يعني أن المحتوى المتطرف غير الغربي غير معروف بما يكفي ولا يحظى بالإدارة الكافية. ونظرًا لأن المقر الرئيسي لأكبر شركات التكنولوجيا في العالم يقع في الغرب، "فقد لا يكون المهندسون والمديرون التنفيذيون المسؤولون عن تصميم المنتجات التكنولوجية على دراية بمحفزات العنف والتمييز في ثقافات أخرى غير ثقافاتهم."[129] يتضح من الوضع في ميانمار ما نراه على المحك بسبب عدم اكتمال تطور تعديل المحتوى غير الغربي وإزالته. وحرض خطاب الكراهية باللغة البورمية على الإنترنت ضد مجتمع الروهينجا على العنف على نطاق واسع. وتصاعدت وتيرة الكراهية العنصرية ضد الأقلية دون رادع إلى حد كبير، لأن فيسبوك استعان بمراجعيْن اثنيْن فقط ناطقيْن باللغة البورمية لمراجعة المحتوى.[130]

وتستطيع منصات التكنولوجيا مواجهة هذا التحيز إذا استفادت من الخبرات الموجودة في المجتمع المدني والأوساط الأكاديمية واستعانت بها في مرحلة تطوير البرمجيات. ينبغي أن تجري الشركات عمليات تدقيق شاملة ومنتظمة لاستخدام الخوارزميات. وينبغي إتاحة نتائج عمليات التدقيق هذه للجمهور لتعزيز المساءلة والشفافية وثقة الجمهور.[131]

يُعد التوسع في قدرات إدارة المحتوى – من الناحية اللغوية والجغرافية للفرق البشرية؛ وفي تطوير أدوات الذكاء الاصطناعي غير الغربية – استثمارًا مكلفًا لوسائل التواصل الاجتماعي وشركات التكنولوجيا. والجهود المبذولة لدعم هذا التوسع الضروري جدًا تُعد فرصة سانحة ينبغي أن تغتنمها شركات التكنولوجيا لتعزيز ريادتها في هذه الصناعة لاسيما في استراتيجيات إدارة المحتوى وإزالته.

التوصية 3: قيام أصحاب المصلحة على الصعيد الوطني والدولي والمبادرات بتمويل منشورات تتناول التكنولوجيا والتحديات والفرص التي يقدمها الذكاء الاصطناعي بلغة واضحة ومفهومة

مثلما بيَّنا في الذكاء الاصطناعي ومكافحة التطرف العنيف: كتاب تمهيدي هناك تربص وضجيج كبير حول الذكاء الاصطناعي. وكثيرون من صانعي السياسات يسيئون فهم حقيقة الذكاء الاصطناعي وما يمكنه أن يفعل. وعلاوة على ذلك، "كشفت المناقشات البرلمانية وجلسات استماع اللجان في المملكة المتحدة في أعقاب فضيحة كامبريدج أناليتيكا 2018، وفي الكونغرس الأمريكي والبرلمان الأوروبي، عن انخفاض صادم في مستويات الفهم والمعرفة الإعلامية بين كبار البرلمانيين وصناع السياسات."[132]

ويمكن أن يؤدي تدني مستوى فهمهم للبيئة الرقمية والإعلامية، فضلاً عن قدرات الذكاء الاصطناعي وحدوده، إلى سياسات تفضي إلى نتائج دون المستوى المنشود. ويحتاج صانعو السياسات إلى تعلم هذا المجال ليتخذوا قرارات سياسية مدروسة ومتوازنة ومستنيرة.

---

126 Crawford, K. (25 يونيو 2016)، 'Artificial Intelligence's White Guy Problem', New York Times. متاح [غير مجاني]: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

127 Angwin, J. et al. (23 مايو 2016)، 'Machine Bias', ProPublica. متاح: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

128 Nieva, R. (1 يوليو 2015)، 'Google apologizes for algorithms mistakenly calling black people "gorillas"', CNET. متاح: https://www.cnet.com/news/google-apologizes-for-algorithm-mistakenly-calling-black-people-gorillas/

129 Elhai, p.221.

130 Stecklow, S. (2018), 'Special Report: Why Facebook is losing the war on hate speech in Myanmar', Reuters. متاح: https://www.reuters.com/article/us-myanmar-facebook-hate-specialreport/special-report-why-facebook-is-losing-the-war-on-hate-speech-in-myanmar-idUSKBN1L01JY

131 Turner Lee, N. (2019), 'Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms', Brookings Institute. متاح: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

132 'Tackling the Information Crisis: A Policy Framework for Media System Resilience', The Report of the LSE Commission on Truth Trust & Democracy, p.38. متاح: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

وينبغي أن تركز المبادرات الوطنية ومتعددة الجنسيات على إنتاج ونشر توجيهات منتظمة عن ماهية الذكاء الاصطناعي، فضلاً عن التحديات والفرص التي يقدمها لعالم السياسة. وينبغي أن يتولى المجتمع المدني والخبراء الأكاديميون المعنيون بما يقع من أضرار عبر الإنترنت وعلى أرض الواقع في سياقات معينة إنتاج كتيبات تمهيدية تفاعلية حول التهديدات الناشئة والمستقبلية. على سبيل المثال، أن يقدم الخبراء المطلعون على الانتخابات القادمة المثيرة للجدل في سياق غير غربي موجزًا لصانعي السياسات حول السياق المعني والعوامل المحفزة للمحتوى الضار وكيف يتحول هذا إلى أضرار على أرض الواقع.

وينبغي كتابة هذه التوجيهات والكتيبات الموجزة بأسلوب واضح ومفهوم، بلا خوض في مصطلحات تقنية أو ضجيج مثير حول الذكاء الاصطناعي. وهكذا، سوف يتمكن صانعو السياسات والعوام على حد سواء من المساهمة في الخطاب العام حول الذكاء الاصطناعي ومكافحة التطرف العنيف.

Global Network
on Extremism & Technology

بيانات الاتصال

لأي أسئلة أو استفسارات، أو للحصول على نسخ أخرى من هذا التقرير، يرجى التواصل مع:

ICSR
King's College London
Strand
London WC2R 2LS
المملكة المتحدة

هاتف: +44 20 7848 2098
بريد إلكتروني: mail@gnet-research.org

تويتر: @GNET_research

هذا التقرير، كغيره من منشورات الشبكة العالمية للتطرف والتكنولوجيا (GNET)، يمكن تنزيله مجانًا من موقع شبكة GNET على الإنترنت www.gnet-research.org.