unesco

UNITED NATIONS OFFICE ON GENOCIDE PREVENTION AND THE RESPONSIBILITY TO PROTECT

وثيقة العمل: معالجة خطاب الكراهية على وسائل التواصل الاجتماعي: التحديات المعاصرة

من إعداد: أنتونيلا بيريني، آن بلوين، ريجا ويس، جوناثان برايت

قام باحثون في معهد أكسفورد للإنترنت بإعداد هذه الوثيقة، بدعم من اليونسكو، كمساهمة في استراتيجية وخطة عمل الأمم المتحدة بشأن خطاب الكراهية، وكذلك في إطار مشروع "#CoronavirusFacts: معالجة وباء التضليل الإعلامي بشأن كوفيد-19 في بيئات معرّضة للصراع" والذي يحظى بتمويل من الاتحاد الأوروبي.
وتشكّل الوثيقة جزءاً من تعاون معهد أكسفورد للإنترنت مع اليونسكو من أجل وضع مجموعة أدوات تحدّد الأساليب والموارد والمشاريع البحثية الحالية والتي تمّ تطويرها لمراقبة وجود خطاب الكراهية على الإنترنت وانتشاره وتأثيره، بالإضافة إلى تقييم القدرات والممارسات للتصدّي له.
نرحّب بكلّ التعليقات التي تخصّ وثيقة العمل هذه والتي من شأنها المساهمة في تعزيز هذه الدراسة الأوسع نطاقاً.

تُعدّ معالجة خطاب الكراهية والتصدّي له مسعى متعدّد المستويات، يشمل معالجة أسبابه الجذرية ودوافعه، ومنع تحوّله إلى عنف، والتعامل مع عواقبه المجتمعية الأوسع نطاقاً. وبُغية تطوير استجابات فعّالة لخطاب الكراهية، بما في ذلك من خلال التعليم، من الضروري مراقبة الظاهرة وتحليلها بشكل أفضل عن طريق الاعتماد على بيانات واضحة وموثوقة. وفي العصر الرقمي، يعني ذلك أيضاً اكتساب فهم أفضل لحدوث خطاب الكراهية على الإنترنت ومدى حدّته وانتشاره.

من المنظور المنهجي، يواجه تحديد خطاب الكراهية على الإنترنت لأغراض البحث، تحدياتٍ عديدةً بما في ذلك التعريفات المستخدمة لتأطير المسألة، والسياقات الاجتماعية والتاريخية، والخصائص اللغوية، ناهيك عن تنوّع مجتمعات الإنترنت وأشكال خطاب الكراهية على الإنترنت (نوع اللغة والصور، الخ). أمّا من المنظور التكنولوجي، فتصعب دراسة خطاب الكراهية على الإنترنت بسبب الموثوقية المتفاوتة لأنظمة الكشف، وعدم شفافية الخوارزميات الخاضعة لحقّ الملكية، بالإضافة إلى صعوبة الوصول إلى البيانات التي تحتفظ بها الشركات وما إلى ذلك. ويُعتبر الوضوح بشأن كيفية التصدّي لهذه التحديات أمراً ضرورياً لاكتساب مزيد من الفهم بشأن كيفية ظهور خطاب الكراهية على الإنترنت وانتشاره، ومن ثمّ لصياغة استجابات فعّالة.

تحدّد استراتيجية وخطة عمل الأمم المتحدة عدداً من المجالات ذات الأولوية لرصد خطاب الكراهية وتحليله، وتنصّ على ضرورة أن تكون الكيانات المعنية التابعة للأمم المتحدة قادرةً على "التعرّف إلى اتجاهات خطاب الكراهية، ورصدها، وجمع البيانات بشأنها وتحليلها".
وعند التركيز على خطاب الكراهية على الإنترنت، يتمّ تشجيع كيانات الأمم المتحدة على الترويج "لمزيد من البحث حول العلاقة بين إساءة استخدام الإنترنت ووسائل التواصل الاجتماعي لنشر خطاب الكراهية والعوامل التي تدفع الأفراد إلى العنف"، وكذلك "تحديد المخاطر والفرص الناشئة في ما يتعلّق بانتشار خطاب الكراهية الذي تطرحه التكنولوجيات الجديدة والمنصّات الرقمية"؛ وأخيراً، "تحديد بروتوكولات العمل التي تأخذ في الاعتبار الأشكال الجديدة لخطاب الكراهية الرقمي".

على مدار العام الماضي، أكّدت جائحة كوفيد-19 أهمية استراتيجية وخطة عمل الأمم المتحدة بشأن خطاب الكراهية، حيث عمّت موجة من خطاب الكراهية جميع أنحاء العالم، ممّا أدى إلى تفاقم التعصّب والتمييز تجاه مجموعات معيّنة وزعزعة استقرار المجتمعات والأنظمة السياسية.

تسعى وثيقة العمل هذه إلى تقديم نظرة عامة عن الجوانب الرئيسية التي يجب أخذها في الاعتبار لمعالجة حدوث خطاب الكراهية على وسائل التواصل الاجتماعي، سواء من خلال أنظمة محدّدة تضعها شركات وسائل التواصل الاجتماعي أو من خلال جهود مضادة وتشريعات أو تدابير وقائية تعليمية. تنقسم الوثيقة إلى ثلاثة أقسام: يركّز القسم الأول على تعريفات خطاب الكراهية والأطر القانونية المرتبطة به؛ فيما يستعرض القسم الثاني أدوات وتقنيات مراقبة خطاب الكراهية على الإنترنت ويناقش مدى انتشار خطاب الكراهية على الإنترنت؛ أمّا القسم الثالث فيتناول بعض الأضرار التي يسبّبها خطاب الكراهية ويناقش التدابير المضادة والوقائية المحتملة.

1
تعريف خطاب الكراهية

بالنسبة إلى معالجة خطاب الكراهية ووضع تشريعات بشأنه، تبدأ الصعوبات بتعريفه. إذ لا يوجد تعريف مقبول دولياً لماهية خطاب الكراهية. بالعكس، يطرح التعريف العديد من المسائل القانونية، على غرار حرية الرأي والتعبير، والتمييز والدعوة أو التحريض على التمييز أو العدائية أو العنف.

يشير **مشروع الخطاب الخطير** لسوزان بينيش[1] إلى وجود صعوبتين رئيسيتين في مصطلح "خطاب الكراهية". أوّلاً، "الكراهية" مصطلح غامض يمكن أن يتّخذ مستويات مختلفة من الشدة ويمكن أن تتبعه عواقب مختلفة: 'هل تعني "الكراهية" في خطاب الكراهية أنّ المتحدث يكره أو يسعى إلى إقناع الآخرين بالكراهية أو يرغب في جعل الناس يشعرون بأنّهم مكروهون؟' [2]ثانياً، يعني "خطاب الكراهية" في جوهره أنّ شخصاً أو مجموعة ما مستهدفون بسبب هويتهم/انتمائهم إلى مجموعة. وهذا يتطلّب أن يحدّد القانون أو التعريف ما إذا كان يعتبر أنّ جميع الهويات والمجموعات تقع تحت طائلة هذا القانون أم لا، وإذا لم يكن الأمر كذلك، ما هو نوع المجموعات المدرجة. يرى مشروع الخطاب الخطير أنّه يمكن إساءة استخدام القوانين الفضفاضة ضد الفئات الضعيفة أو المعارضة السياسية والمدنية، ممّا يؤدي في بعض الأحيان إلى إلحاق الضرر بنفس المجموعات التي ترمي قوانين خطاب الكراهية إلى حمايتها. ومع ذلك، يمكن القول إنّ اعتماد تعريف يركّز بشكل ضيّق للغاية على مجموعات وهويات محدّدة يمكن أن يؤدي إلى استبعاد قانوني أو افتقار للأدوات القانونية لمعالجة المشكلة.

رغم أنّ نطاق وثيقة العمل هذه لا يسمح بدراسة هذه التحديات بالتفصيل، إلاّ أنّ استعراض القوانين الدولية والوطنية حول العالم يُبيّن التعقيدات والتفسيرات المختلفة التي يمكن تطبيقها على خطاب الكراهية.

على الصعيد العالمي، إلى جانب الإعلان العالمي لحقوق الإنسان وهو صكّ غير ملزم، يقوم العهد الدولي الخاص بالحقوق المدنية والسياسية بمتابعة الحقّ في حرية التعبير (المادة 19) مع حظر أي دعوة إلى الكراهية تشكل تحريضاً على التمييز أو العداوة أو العنف (المادة 20). كما تقيّد المادتان 19 و20 حرية التعبير، شريطة أن تكون هذه القيود **محدّدة بموجب القانون** وأن تكون ضرورية: (أ) **لاحترام حقوق الآخرين أو سمعتهم؛** (ب) **لحماية الأمن القومي أو النظام العام، أو الصحة العامة أو الآداب العامة.**

واستكمالاً لهذه المبادئ، تقترح خطة عمل الرباط **"اختبار الحدّ المسموح به من ستة نقاط"** لتبرير القيود المفروضة على حرية **التعبير**، مع أخذ **السياق الاجتماعي والسياسي، ووضع المخاطِب، ونية التحريض على العداوة، ومحتوى الخطاب، ومدى انتشار الخطاب، واحتمالية إحداث الضرر** في الاعتبار.

ومن بين الصكوك البارزة أيضاً في مجال خطاب الكراهية، نجد الاتفاقية الدولية للقضاء على جميع أشكال التمييز العنصري التي تضمّ بنداً أكثر صرامة من المادة 20 التابعة للعهد الدولي الخاص بالحقوق المدنية والسياسية بما أنّه لا يتطلّب النية أو "الدعوة إلى الكراهية" ويشمل النشر في قائمة الممارسات التي يعاقب عليها القانون. ومن بين الصكوك ذات الصلة بهذا المجال، نجد اتفاقية منع جريمة الإبادة الجماعية والمعاقبة عليها واتفاقية القضاء على جميع أشكال التمييز ضد المرأة.

وضعت منظمة حرية التعبير ARTICLE 19 مبادئ كامدن حول حرية التعبير والمساواة على أساس المناقشات مع مسؤولي الأمم المتحدة وخبراء من الأوساط الأكاديمية والمجتمع المدني. وتوفّر هذه المبادئ توجيهاتٍ تفسيريةً لمواد العهد الدولي الخاص بالحقوق المدنية والسياسية وتسعى إلى ردع الجهات الفاعلة عن إساءة استخدام المادة 20 من خلال تحديد المسائل المتعلقة بـ "التحريض"، بالإضافة إلى ما يشكّل "تمييزاً" و "عدائيةً" و "عنفاً".

[1] مشروع الخطاب الخطير، 2021، https://dangerousspeech.org/

[2] بينيش، سوزان (2021) الخطاب الخطير: دليل عملي. مشروع الخطاب الخطير، 2021، ص 7، https://dangerousspeech.org/

عند ترجمة القانون والمبادئ الدولية إلى قانون وطني، يكون لكلّ بلد مقاربة مختلفة بعض الشيء لكيفية تعريفه لخطاب الكراهية من حيث طريقة التعبير عنه، والضحايا المحتملين، ونوع الضرر الذي يجب أن يحدث ليُعتبر الخطاب خطاب كراهية. يُشكّل عدم وجود تعريف موحّد أحد التحديات الرئيسية عندما يتعلّق الأمر بمكافحة خطاب الكراهية على الإنترنت والذي لا يقتصر بالضرورة على الحدود الوطنية.

كما تؤدي تعريفات خطاب الكراهية دوراً مهماً بالنسبة إلى جهود البحث والدفاع، لاسيما عند تحديد عواقبه المجتمعية. وقد تقع الأضرار الناجمة عن خطاب الكراهية على مستوى الأفراد (على شكل ضرر نفسي) والمجموعات والجماعات والمجتمع (على شكل تآكل للحقوق والمنافع العامة). وبما أنّ خطاب الكراهية يستهدف الأشخاص على أساس سمات مجموعة ما، فإنّ للتحليل على مستوى الضرر المجتمعي أهميةً بالغةً. وتتوزّع الأضرار الناجمة عن خطاب الكراهية بشكل غير متساو بين عموم السكان وتتحمّل الفئات المهمشة العبء الأكبر. كما تتراكم هذه الأضرار بالنسبة إلى الأشخاص الذين يعانون منها، علماً بأنّ التجارب السابقة بيّنت أنّ خطاب الكراهية يشكّل متغيّراً رئيسياً في تقدير الضرر الناجم عن الاستهداف بخطاب الكراهية.[3]

خطاب الكراهية على الإنترنت

لا يختلف خطاب الكراهية على الإنترنت في جوهره عن خطاب الكراهية خارج الفضاء الإلكتروني. ومع ذلك، فهو يختلف عنه في طبيعة التفاعلات عند حدوثه، وكذلك في استخدام وانتشار كلمات محدّدة واتهامات ونظريات المؤامرة التي يمكن أن تتطوّر وتبلغ ذروتها قبل أن تتلاشى بسرعة كبيرة. ويمكن أن تنتشر الرسائل المحرّضة على الكراهية في غضون ساعات أو حتّى دقائق.

يشير تقرير اليونسكو "التصدي لخطاب الكراهية على الإنترنت" لعام 2015 إلى كيفية إعداد خطاب الكراهية على الإنترنت ونشره بتكلفة منخفضة، بحيث لا يخضع لعملية تحرير مثل الأعمال المكتوبة الأخرى. كما يشهد مستويات متباينة من التعرّض، وذلك بحسب شعبية المنشور، ويمكن نشره عبر البلدان بدون الحاجة بأن تكون خوادم المنصات ومقراتها في في نفس البلد الذي يتواجد فيه المستخدم والجمهور المستهدف. كذلك، يمكن أن يكون خطاب الكراهية على الإنترنت متاحاً لفترة أطول وأن يمرّ بموجات من الشعبية، أو أن يتّصل بشبكات جديدة أو أن يظهر مجدداً، بالإضافة إلى الحفاظ على سرية هوية صاحبه.

ونتيجة لذلك، شكّلت مسألة هوية من يشرف على الفضاءات الإلكترونية وشروط وتوقيت إزالة المحتوى، موضوع نقاش واسع النطاق.

ويتجسّد هذا النقاش في قوانين مثل **قانون "NetzDG" الألماني** (قانون إنفاذ الشبكة) الذي تمّ تقديمه في عام 2017 والذي يطلب من منصات الوسائط الاجتماعية التي تضمّ أكثر من مليوني مستخدم تطبيق إجراءات شفّافة لتعديل المحتوى غير القانوني (بما في ذلك خطاب الكراهية) وإزالة المحتوى الذي تمّ تحديده على أنّه غير قانوني خلال فترة زمنية مدّتها 24 ساعة ورفع تقارير منتظمة بشأن التدابير المتخذة. وقد تعرّض هذا القانون لانتقادات شديدة لأنّه دفع المنصات إلى تأدية دور "رقابة مخصخصة" بشأن قرارات ينبغي أن تتّخذها المحاكم. كما تمّ التحذير بأنّ المهل الزمنية والغرامات قد تؤدّي إلى "إزالة مفرطة" للمحتوى من المنصات لتجنّب خطر العقوبات المشدّدة. في عام 2020، تمّ تعديل القانون ليطلب من منصات وسائل التواصل الاجتماعي إحالة المحتوى المحدّد على أنّه غير قانوني إلى مكتب الشرطة الجنائية الفيدرالية. وفي الوقت نفسه تقريباً، أجريَ تعديل آخر ساهم في تعزيز حقوق المستخدمين من خلال مطالبة المنصات بجعل الإبلاغ عن المحتوى غير القانوني أكثر سهولة، بالإضافة إلى السماح بالطعن في قرار حذف أو عدم حذف منشور ما.

وفي هذا السياق، غالباً ما يكون تطوير قوانين تهدف إلى التصدّي لخطاب الكراهية على الإنترنت وخارجه، محفوفاً بعمليات مراجعة معقدة تتعلق بتحديات التعريف وبمهمة احترام حرية التعبير في إطار القانون على حد سواء. ونظراً لهذه التحديات، يجب أيضاً استخدام أساليب تتجاوز التدابير القانونية لمعالجة خطاب الكراهية.

[3] كاثرين جيلبر ولوك ماكنمارا (2016) إثبات أضرار خطاب الكراهية، الهويات الاجتماعية، 22:3

2
أدوات وتقنيات قياس خطاب الكراهية ورصده

تختلف السياسات والأدوات المتعلقة بالكشف عن خطاب الكراهية على الإنترنت ورصده وتعديله، بحسب السياقات والجهات الفاعلة والمنصات.

يمكن توزيع طرق الكشف إلى فئتين: جهود أكثر شموليةً اعتمدت في البداية على فلترة الكلمات المفتاحية وطرق حشد المصادر، والجهود التي تعتمد على مشرفين بشريين يراجعون المحتوى المبلّغ عنه كخطاب كراهية من قبل المستخدمين ويقرّرون ما إذا كان يندرج ضمن ذلك الإطار. فيما تتمتّع المقاربات اليدوية بميزة واضحة تتمثّل في التقاط السياق والتجاوب السريع مع التطورات الجديدة، تتطلّب العملية يداً عاملة كثيفة وتستهلك قدراً كبيراً من الوقت والموارد، ممّا يحدّ من قابلية التوسع وتوفير حلول سريعة. ونتيجة لهذه التحديات وللحجم المتزايد للمحتوى المنتج على وسائل التواصل الاجتماعي والتقدم المحرز في التعلم الآلي وفي تقنيات معالجة اللغة الطبيعية، قامت المنصات والباحثون بتطوير حلول آلية للكشف وباتوا يعتمدون عليها بشكل متزايد. تستخدم العديد من المبادرات الجديدة مجموعةً متنوعةً من الأساليب. وتشمل بعض المصطلحات الأساسية المتعلقة بهذه المنهجيات:

- **التعلم الآلي:** وهي تقنيات تستخدم خوارزميات الكمبيوتر التي يمكن أن تتحسّن آلياً من خلال التجربة واستخدام البيانات.
- **معالجة اللغة الطبيعية:** وهي تقنيات تعالج وتحلّل كمياتٍ كبيرةً من بيانات اللغة الطبيعية.
- **المقاربات القائمة على الكلمات المفتاحية:** وهي طرق تستخدم الأنطولوجيا أو القاموس، وتحدّد النص الذي يحتوي على كلمات مفتاحية يحتمل أن تكون محرّضة على الكراهية.
- **الدلالات التوزيعية:** وهي طرق لتحديد وتصنيف أوجه التشابه بين الكلمات والعبارات والجمل بناء على كيفية توزيعها في عينات كبيرة من البيانات.
- **تحليل المشاعر:** وهي طرق لشرح نوع المواقف التي يتمّ نقلها بشأن بموضوع ما في نص معيّن.
- **البيانات الوصفية للمصدر:** تقوم بعض الطرق بإبلاغ النماذج من خلال المعلومات الوصفية للبيانات، على غرار البيانات المتعلقة بالمستخدمين المرتبطين بالرسائل، بما في ذلك الميزات القائمة على الشبكة مثل عدد المتابعين.
- **التعلّم العميق:** وهي فئة من خوارزميات التعلّم الآلي تستخدم طبقاتٍ متعددةً لاستخراج ميزات ذات مستوى أعلى من المدخلات الأولية بشكل تدريجي.

لقد انتقلت شركات وسائل التواصل الاجتماعي إلى حدّ كبير من مجرّد الردّ على المنشورات التي أبلغ عنها المستخدمون كخطاب كراهية إلى اكتشاف ومعالجة هذا المحتوى بشكل استباقي من خلال أنظمتها الآلية قبل أن يراها المستخدمون. وعلى الرغم من ضرورة معالجة خطاب الكراهية على نطاق واسع، إلاّ أنّ هذه الأساليب تنطوي كذلك على تعقيدات: يؤدي الكشف الآلي لخطاب الكراهية حتماً إلى ارتكاب أخطاء وقد يقود إلى إزالة محتوى لا يحرّض على الكراهية. وقد تؤدي الإزالة المفرطة للمحتوى إلى قمع ممارسة الحقوق وتقويض حرية التعبير.

وبهدف تحسين أنشطة الرصد الخاصة بها، يتمّ باستمرار تطوير أدوات الكشف عن الخطاب الكراهية. على سبيل المثال، Perspective API[4] هي أداة مفتوحة المصدر من Jigsaw (حاضنة داخل جوجل) وفريق تكنولوجيا مكافحة إساءة الاستخدام التابع لجوجل استخدمتها المؤسسات الإخبارية ومنتجات جوجل. وتستخدم هذه الأداة التعلّم الآلي لتسجيل العبارات بناء على سُمّيتها المحتملة في محادثة ما. وهي متاحة بسبع لغات (الإنجليزية والفرنسية والألمانية والإيطالية والبرتغالية والإسبانية والروسية). وصرّح فيسبوك أنّ أحدث إصدار من أداته الرامية إلى مراقبة وكشف خطاب الكراهية على منصاته قد حسّن الفهم الدلالي للغة وفهم المحتوى، بما في ذلك تحليل الصور والتعليقات والعناصر الأخرى.[5]

كما عمل الباحثون ومنظمات المجتمع المدني على تطوير أدوات للكشف عن خطاب الكراهية. تشمل بعض الأمثلة ما يلي:

[4] يتوفر مزيد من المعلومات حول Perspective API على https://www.perspectiveapi.com/
[5] المصدر: https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/

انتشار خطاب الكراهية على منصات التواصل الاجتماعي

- كانت المنصة الكينية **Umati** من أوائل الشركات التي تراقب يدوياً المنشورات على الإنترنت المكتوبة باللغات المستخدمة في كينيا.
- طوّر **Hate Sonar** Davidson et al. (2017) باستخدام مقاربة انحدار لوجستي مُدرَّبة على البيانات من منتديات الويب و تويتر.
- طوّر ADL و D-Lab من جامعة بيركلي **مؤشّر الكراهية** على الإنترنت، المصمّم لتحويل الفهم البشري لخطاب الكراهية عبر التعلّم الآلي إلى أداة قابلة للتطوير يمكن نشرها على محتوى الإنترنت لاكتشاف نطاق وانتشار خطاب الكراهية على الإنترنت.
- طوّر فريق مشروع التدابير والإجراءات المضادة في معهد آلان تورنغ أداة تستخدم أساليب التعلم العميق **للكشف عن التحيّز ضدّ الآسيويين الشرقيين** على وسائل التواصل الاجتماعي.
- طوّر(Moon et al. (2020 أداةً للكشف عن خطاب الكراهية ضد الكوريين، حيث تمّ تدريب النموذج على علامة "التحيز" بالإضافة إلى "الكراهية".
- يكشف **مقياس الكراهية** عن خطاب الكراهية ضد المسلمين باستخدام التعلّم الآلي وتقنيات معالجة اللغة الطبيعية. إنّ المنصة متاحة باللغات الإنجليزية والفرنسية والإيطالية.
- يقوم **COSMOS** بجمع وتحليل البيانات من تويتر في الوقت الفعلي من خلال مواصفات الكلمات المفتاحية، باستخدام تحليل المشاعر ومعالجة اللغة الطبيعية.
- يقوم **MANDOLA** بالكشف عن المحتوى المحرّض على الكراهية من خلال مزيج من تحليل المشاعر ومعالجة اللغة الطبيعية والتعلّم الآلي والتعلّم العميق.

وتجدر الإشارة إلى أنّ رصد خطاب الكراهية على الإنترنت يعتمد على إمكانية الوصول إلى البيانات، لاسيما من منصات وسائل التواصل الاجتماعي. علاوة على ذلك، فإنّ العديد من الأدوات القائمة أحادية اللغة، وغالباً ما تقتصر على اللغة الإنجليزية، وهناك حاجة لإجراء مزيد من الأبحاث حول أداء طرق الكشف متعدّدة اللغات. بالإضافة إلى ذلك، صبّت الأغلبية الساحقة من البحوث وعمليات رصد خطاب الكراهية على منصات وسائل التواصل الاجتماعي، تركيزها على الولايات المتحدة وأوروبا، ممّا أدى إلى فجوة ليس في الأدوات والبيانات فحسب، لا بل أيضاً في فهم نطاق وديناميكيات انتشار خطاب الكراهية في مناطق أخرى. لذا أصبح سدّ هذه الفجوة ضرورةً ملحةً نظراً للطابع السياقي لخطاب الكراهية.

باستخدام أدوات الكشف الآلي القائمة على الأساليب المتاحة اليوم، أبلغ تويتر وفيسبوك وانستغرام ويوتيوب بشكل متزايد عن محتويات تمّ الإبلاغ عنها و/أو حذفها. بين كانون الثاني/يناير وآذار/مارس 2021، أزال يوتيوب 85247 مقطع فيديو انتهك سياسة خطاب الكراهية. ويظهر تقريراهما السابقان أرقاماً مماثلة. وخلال الربع نفسه، أبلغ فيسبوك عن التعامل مع إجمالي 25.2 مليون جزء محتوى، بينما أبلغ انستغرام عن 6.3 مليون جزء محتوى. وفقاً لتقرير الشفافية الأخير لتويتر، قامت الشركة بإزالة 1628281 جزء محتوى تمّ اعتبار أنّها تنتهك سياسة خطاب الكراهية بين تموز/يوليو وكانون الأول/ديسمبر 2020.

على منصات وسائل التواصل الاجتماعي، يتمّ تحديد انتشار خطاب الكراهية من خلال عينة من المحتوى الذي يشاهده المستخدمون. بتعبير آخر، يتمّ فقط التقاط (تقدير لـ) ما يتبقّى من خطاب الكراهية على المنصة بخلاف ما كشفت عنه الشركة بالفعل إزالته بشكل استباقي. حتى الآن، يُعدّ فيسبوك المنصة الوحيدة التي تبلغ عن مقاييس الانتشار. إذ أفادت الشركة أنّه بين كانون الثاني/يناير 2021 وآذار/مارس 2021 كان هناك انتشار لخطاب الكراهية بنسبة تتراوح بين 0.05% و0.06%، ممّا يظهر انخفاضاً طفيفاً مقارنة بتقريريها السابقين. وتشير بعض الدراسات إلى أنّ انتشار خطاب الكراهية على المنصات الرئيسية، مثل تويتر وويكيبيديا، يمثّل أقلّ من 1% من إجمالي المحتوى، بينما تتراوح نسبة المحتوى المسيء في المنصات البديلة الأكثر تخصصاً، مثل Gab و4chan، بين 5 [6]و8%. لا تزال أدلة انتشار خطاب الكراهية على منصات وسائل التواصل الاجتماعي غير كاملة، ويعزى ذلك جزئياً إلى انعدام الشفافية والوصول إلى البيانات من جانب المنصات.

[6] Zannettou et al., 2018; Mathew et al., 2018; Hine et al., 2017, Vidgen et al, 2019.

[7] منذ عام 2013، رفعت مبادرة اليونسكو لتدريب القضاة من قدرات الجهات القضائية الفاعلة فيما يخصّ المعايير الدولية والإقليمية المتعلقة بحرية التعبير والنفاذ إلى المعلومة وسلامة الصحفيين في مناطق في جميع أنحاء العالم، باعتبار أنّ مسألة تحديد أفضل السبل التعامل مع مسائل خطاب الكراهية هو أحد الاهتمامات الرئيسية للعديد من العاملين في المجال القضائي. وقد تمّ تدريب أكثر من 23000 فاعل قضائي على هذه المسائل، لا سيما من خلال سلسلة من الدورات الإلكترونية المفتوحة الحاشدة والتدريب الميداني وورش العمل، وإصدار عدد من مجموعات الأدوات والمبادئ التوجيهية

3
مكافحة خطاب الكراهية على الإنترنت

تجدر الإشارة أولاً إلى أنّ مواجهة خطاب الكراهية، وبالتالي اختيار الأدوات والاستراتيجيات المناسبة والجهود الوقائية، أمر معقّد بسبب عوامل عدّة. هناك إجماع ضعيف في إجابات الجهات الفاعلة المختلفة على الأسئلة الرئيسية في مجموعة من السياقات. كيف يحدث خطاب الكراهية ضرراً، ومتى يكون الضرر شديداً بما فيه الكفاية لتبرير تنظيم الخطاب؟ وبشكل أكثر تفصيلاً، ما هي أنواع الضرر المرتبطة بسلوكيات خطاب الكراهية، التي تستدعي وجود أنظمة تتماشى مع القانون الدولي لحقوق الإنسان ومعايير حرية التعبير؟

كما تضيف هندسة الإنترنت تحدياتٍ مختلفةً لمواجهة خطاب الكراهية. وتشمل هذه التحديات الاستمرارية، التجوال، عدم كشف الهوية، والطابع متعدّد الاختصاصات للمحتوى على الإنترنت، بالإضافة إلى مجموعة واسعة من هندسات المنصات، ونظام غير متجانس لحوكمة الإنترنت يشارك فيها أصحاب مصالح متعددون.

رغم هذه التحديات، فإنّ العديد من المجموعات والأفراد يشاركون بطرق مختلفة في مكافحة خطاب الكراهية على الإنترنت وتعزيز مرونة مستخدمي الإنترنت ضده بشكل وقائي.

اللجوء إلى قضاء الدولة

يتمثّل السبيل البارز في التصدي لخطاب الكراهية في اللجوء إلى القضاء. رغم أنّ المواقف من خطاب الكراهية وخطاب الكراهية على الإنترنت تختلف بين المناطق وتواصل التطور مع فهم المسألة بشكل أفضل، إلّا أنّ هناك عدداً من المبادئ الدولية والاتفاقيات الإقليمية والقوانين على صعيد الدول وأمثلة على الاجتهاد القضائي تتماشى مع المعايير الدولية لحقوق الإنسان التي تتضمّن بنوداً ذات صلة بخطاب الكراهية على الإنترنت وخارج الفضاء الإلكتروني، كما هو موضّح في بداية هذه الوثيقة.

إلا أنّ المسائل المتعلقة بالاستجابات القانونية البحتة على خطاب الكراهية سرعان ما تنشأ على الإنترنت. ويشمل ذلك مخاوف بشأن توازن الحقوق، وإمكانية قيام الجهات الفاعلة القوية بإساءة استخدام القيود المفروضة على الحقوق، والاعتماد على عتبات لمنع التحريض على العنف إلى جانب علاقة غير مفهومة بشكل جيّد بين خطاب الكراهية والعنف خارج الفضاء الإلكتروني. والأهمّ من ذلك وبهدف التصدي لخطاب الكراهية على الإنترنت، تتمثّل إحدى المسائل الرئيسية للجوء إلى القضاء في تقييد سلطة الدول الفردية على الفضاءات الرقمية على الإنترنت. فلا يمكن أن تعتمد المعالجة الفعالة لخطاب الكراهية على الإنترنت على اللجوء إلى قضاء الوطني فقط.[7]

استجابات من جانب شركات التكنولوجيا

في عام 2016، وافقت مجموعة من شركات التكنولوجيا الكبرى على مدونة قواعد السلوك الخاصة بالمفوضية الأوروبية بشأن مكافحة خطاب الكراهية غير القانوني على الإنترنت والتي تطلب من هذه الشركات استعراض خطاب الكراهية في غضون يوم واحد من تلقي الابلاغ. وتمثّل هذه المقاربة تحدياً بسبب التباين الكبير من حيث الخدمة والتعريفات التشغيلية لخطاب الكراهية، لكنّها تمثل جهداً كبيراً لتعزيز التعاون والترابط بين المقاربات القانونية وغير القانونية في فضاء خطاب الكراهية.

في عام 2021، أبلغ كلّ من يوتيوب وفيسبوك عن زيادة في المحتوى الذي تمّ العثور عليه والإبلاغ عنه من قبل كلّ منهما، إلى جانب نسبة أكبر من المحتوى الذي تمّ الإبلاغ عنه من قبل الشركة مقارنة بالمحتوى الذي تمّ الإبلاغ عنه من قبل المستخدمين. ويعود ذلك إلى الاستخدام المتزايد لأنظمة الكشف الآلي. ومع ذلك، فقد اقترن هذا الاتجاه بارتفاع في المحتوى المعاد نشره مقارنةً بالفترات السابقة المبلّغ عنها. فبين كانون

الثاني/يناير وآذار/مارس 2021، أعاد فيسبوك نشر 408700 جزء محتوى وأعاد انستغرام نشر 43700 جزء محتوى. وفيما تشير التقارير إلى أنّ المنصات تتعامل بشكل متزايد مع المحتوى الذي يحرّض على الكراهية، إلاّ أنّنا نجهل ما إذا كان السبب وراء ذلك يكمن في زيادة مصاحبة في نسبة إساءة الاستخدام أو زيادة صرامة سياسات المنصات أو زيادة الإيجابيات الزائفة.

تنشأ شركات وسائل التواصل الاجتماعي ضمن الولايات القضائية الوطنية، وبذلك تتأثر بشكل مباشر بالقوانين الوطنية وعادة ما تكون أكثر استجابةً لطلبات احتواء خطاب الكراهية نتيجة لذلك. ومع ذلك، ليست منصات وسائل التواصل الاجتماعي ملزمةً بمبدأ الاختصاص المكاني وبالتالي يتمّ الاعتماد عليها فقط لتنفيذ شروط الخدمة الخاصة بها والتي قد تكون أو لا تكون أكثر صرامةً من المعايير المنصوص عليها في الاتفاقيات الدولية الموضحة في قسم سابق. تشمل الإجراءات التي تتّخذها منصات وسائل التواصل الاجتماعي إزالة المواد عندما يُنظر إليها كخطاب كراهية، بالإضافة إلى إرسال تحذيرات إلى المستخدمين الذين ينشرون خطاب الكراهية أو تقييد نشاطهم على المنصة أو حظرهم. وتتطوّر معايير المجتمع هذه باستمرار،لاسيّما في ما يتعلّق بمدى اعتمادها على أساليب الإشراف الآلي مقارنةً بأساليب الإشراف البشري.
وعلى ضوء هذه التحديات، فإنّ حركةً يشارك فيها أصحاب مصالح متعددون وتدعو إلى مزيد من الشفافية من طرف شركات الإنترنت كوسيلة لتعزيز مساءلتها، قد اكتسبت زخماً متزايداً في السنوات الأخيرة. وقد شمل ذلك تدابير قانونية وتنظيمية مقترحة في 30 دولة ومنطقة على الأقلّ، بما في ذلك من خلال قانون الخدمات الرقمية الأوروبية الذي لا يزال قيد الإعداد حالياً. كما اتخذت الشركات خطواتٍ لتكون أكثر شفافيةً. ففي عام 2021، قامت Access Now بفهرسة أكثر من 70 شركة تصدر تقارير شفافية [8]منتظمة ، على الرغم من الحاجة إلى المزيد.

يقدّم العرض الموجز لليونسكو بعنوان "ترك الشمس تشرق: الشفافية والمساءلة في العصر الرقمي" تعزيز الشفافية كطريقة ثالثة بين إفراط الدولة في تنظيم للمحتوى، ممّا أدّى إلى قيود غير متناسبة على حقوق الإنسان، ومقاربة عدم التدخل التي فشلت في معالجة المحتوى الذي يطرح إشكالاً بشكل فعال مثل خطاب الكراهية والمعلومات المضللة. ويقدّم الموجز مجموعةً تتكوّن من 26 مبدأ رفيع المستوى تشمل المسائل المتعلقة بالمحتوى والعملية، والعناية الواجبة والاصلاح، والتمكين، والأبعاد التجارية، وجمع البيانات الشخصية واستخدامها، والوصول إلى البيانات.

الإستجابات خارج اختصاص المحاكم والتدخلات الوقائية

إنّ الاستجابات الأخرى الآتية من خارج اختصاص المحاكم هي نتيجة جهود البحث والدعوة التي يبذلها المجتمع المدني أو قد تركّز على التدابير الوقائية التي تعزّز مرونة مستخدمي الإنترنت في مواجهة خطاب الكراهية. وتشمل هذه الإجابات مبادرات تستهدف بشكل مباشر أسباب خطاب الكراهية على الإنترنت وعواقبه بما في ذلك من خلال التعليم، بالإضافة إلى مبادرات تدعو إلى تنفيذ تدابير قانونية وتقنية أفضل.

وتأتي المبادرات القائمة على التعليم في صلب هذه الجهود، وغالباً ما تركّز على الوقاية على المدى الطويل. ويمكن أن تعمل التدخّلات التعليمية على زيادة الوعي بالعواقب الضارة لخطاب الكراهية ومعالجة أسبابه الجذرية والتنبيه الفعّال لتقنيات التلاعب والخطاب المستخدمة لنشر الكراهية على الإنترنت وخارج الفضاء الإلكتروني. وقد تمّ، وعلى وجه الخصوص، تطوير برامج للتثقيف الإعلامي والمعلوماتي وتنفيذها في جميع أنحاء العالم بهدف تزويد مستخدمي الإنترنت بالمهارات اللازمة لفحص معلومات المحتوى على الإنترنت بشكل نقدي وتحديد المحتوى المزعج والمحرّض على الكراهية والمعلومات المضللة. وقد تمّ أيضاً بذل جهود "الخطاب المضاد" التي تهدف إلى معالجة خطاب الكراهية بخطابات مضادة إيجابية، مثل مبادرة الشجاعة المدنية على الإنترنت التي يقودها فيسبوك، والتي أجريت في ألمانيا والمملكة المتحدة وفرنسا في عام 2017. وتركّز مبادرات المجتمع المدني الأخرى على الدعوة للتغيير من جانب المنصات. ففي تموز/يوليو 2020، جمعت حملة Stop Hate for Profit تحالفاً يضم أكثر من 1200 شركة من جميع أنحاء العالم ودعت إلى مقاطعة الإعلانات في المنصات الرئيسية مطالبين إياها بمراقبة خطاب الكراهية وإيقاف الإعلانات مؤقتاً على الحسابات التي تروّج للتمييز ضد مجموعات معيّنة. والتحقت هذه الحملة بعدد متزايد من الأصوات التي دعت إلى معالجة خطاب الكراهية على الإنترنت، لاسيّما على ضوء الاستهداف المكثف للفئات المهمشة خلال جائحة كوفيد-19، ممّا دفع العديد من شركات وسائل التواصل الاجتماعي إلى إدخال تغييرات على المبادئ التوجيهية لمجتمعها.

[8] https://www.accessnow.org/transparency-reporting-index/

التوصيات

لتعزيز عملية صنع السياسات القائمة على الأدلة للحدّ من خطاب الكراهية على الإنترنت، ولمنع ترجمة خطاب الكراهية إلى عنف مع حماية حرية التعبير أيضاً، من المهمّ التعرّف إلى اتجاهات خطاب الكراهية ورصدها وجمع البيانات عنها وتحليلها بهدف تحديد الاستراتيجيات المناسبة لمواجهتها. وترمي التوصيات الواردة أدناه إلى تحديد الإجراءات الرئيسية للتصدي للتحديات الجديدة في انتشار خطاب الكراهية الناشئ وبخاصة معالجة عواقبه خارج الفضاء الإلكتروني على السلام والاستقرار وتمتّع الجميع بحقوق الإنسان.

1. **تعزيز التعريفات الشاملة لخطاب الكراهية التي تحترم حرية التعبير**
   - التأكد بأنّ التعريفات تتماشى مع المعايير الدولية، لاسيّما على النحو المنصوص عليها في العهد الدولي الخاص بالحقوق المدنية والسياسية وخطة عمل الرباط.
2. **بناء تحالفات ذات أصحاب مصلحة متعددين**
   - تشجيع تبادل البيانات والخبرات بين منظمات حقوق الإنسان ووسطاء الإنترنت والجمهور
   - تمكين أصحاب المصلحة، لاسيّما الجماعات المحلية، من رصد وكشف خطاب الكراهية على وسائل التواصل الاجتماعي المصممة وفقاً لسياقهم ولغاتهم.
   - عقد حوارات بين أصحاب المصلحة المتعددين حول اتجاهات خطاب الكراهية وحدوثه وكيفية مواجهته.
   - دعوة المنصات لتطوير التعريفات والإجراءات التشغيلية بالتعاون مع مجموعات الخبراء والجمهور، والتي يجب أن تغطّي المناطق التي تقع خارج أمريكا الشمالية وأوروبا الغربية لتشمل المزيد من البلدان حول العالم.
3. **جمع البيانات وتشجيع ممارسات البيانات المفتوحة التي يتمّ فيها جمع البيانات بالفعل، مع احترام حماية البيانات الشخصية**
   - جمع البيانات النوعية مع الأفراد المستهدفين بخطاب الكراهية لفهم نطاق وطبيعة الأضرار بشكل أفضل
   - الدعوة إلى أن تقوم شركات منصات الإنترنت بتحسين ممارسات الشفافية الخاصة بها، بما في ذلك عن طريق الافصاح العلني عن البيانات حول شكاوى خطاب الكراهية والحلول ذات الصلة، بالإضافة إلى دقة وسير عمل أنظمة إدارة المحتوى الخاصة بها، لاسيّما لأغراض البحث.
   - دعم تطوير أدوات ومنهجيات معقولة التكلفة ومتاحة وسهلة الاستعمال، يمكن استخدامها لرصد خطاب الكراهية والكشف عنه عبر سياقات متعددة اللغات والثقافات ضمن مهلة زمنية تسمح باتخاذ إجراءات مضادة.
4. **تشجيع المنصات على تقديم خيارات إصلاحية فعّالة للأشخاص الذين تمّت إزالة محتواهم**
   - تسهيل التعاون بين شركات وسائل التواصل الاجتماعي ومجموعات المجتمع المدني الذي يركّز على الحقوق الرقمية للتأكد من أنّ عمليات الإشراف على المحتوى وحذفه تتوافق مع احتياجات المجتمع.
5. **تطوير الإلمام بخطاب الكراهية، والتثقيف الإعلامي والمعلوماتي، فضلاً عن المهارات الرقمية من خلال برامج التعليم**
   - توفير التمويل والموارد لتطوير البرامج التعليمية التي تعزّز المرونة ضد خطاب الكراهية، بالاسترشاد بالاتجاهات الحالية لخطاب الكراهية والاستجابة للتحديات ذات الصلة. ويتطلب ذلك تعاوناً وثيقاً بين شركات وسائل التواصل الاجتماعي ومعاهد البحث وأصحاب المصلحة في مجال التعليم.
   - إعطاء الأولوية للمقاربات التعليمية الوقائية التي تنبّه إلى الآثار الضارة لخطاب الكراهية على الإنترنت وتعزّز التثقيف الإعلامي والمعلوماتي إلى جانب جهود التخفيف والتصدّي.
   - إنشاء ودعم الشراكات بين المؤسسات التعليمية وشركات وسائل التواصل الاجتماعي لزيادة إمكانية الوصول إلى المعلومات والموارد للتصدي لخطاب الكراهية على منصات وسائل التواصل الاجتماعي، وذلك عن طريق حملات نشر محدّدة الهدف أو إعادة توجيه المستخدمين نحو الموارد الخارجية.
6. **دعم المنظمات النشطة في فضاء خطاب الكراهية على الإنترنت -**
   - التأكّد من أن الموارد الكافية متوفّرة للمنظمات المتخصصة في رصد خطاب الكراهية والتصدي له، لاسيّما المنظمات الأفضل تجهيزاً لأخذ السياقات المحلية في الاعتبار، وتزويدها بالدعم اللازم.

هذه الورقة جزء من مجموعة أوراق المناقشة ، التي تمّ إعدادها بتكليف من اليونسكو ومكتب الأمم المتحدة للمستشار الخاص المعني بمنع الإبادة الجماعية(OSAPG) ، وتعتبر الأوراق مساهمة مباشرة في استراتيجية وخطة عمل الأمم المتحدة. في سياق منتدى أصحاب المصلحة المتعددين والمؤتمر الوزاري المعني بمعالجة خطاب الكراهية من خلال التعليم في سبتمبر وأكتوبر 2021.

وقد سلط تفشي جائحة COVID-19 الضوء على أهمية استراتيجية وخطة عمل الأمم المتحدة ، مما أدى إلى ظهور موجة من خطاب الكراهية في جميع أنحاء العالم - مما يؤدي إلى تفاقم التعصب والتمييز تجاه مجموعات معينة وزعزعة استقرار المجتمعات والأنظمة السياسية. تسعى أوراق المناقشة إلى تفريغ القضايا الرئيسية المتعلقة بهذا التحدي العالمي واقتراح الاستجابات والتوصيات الممكنة.

تمّ إجراء هذه الورقة بتكليف من قسم حرية التعبير وسلامة الصحفيين في اليونسكو كجزء من مشروع "#CoronavirusFacts" ، معالجة "التطهير بشأن COVID-19" في البيئات المعرضة للصراع بتمويل من الاتحاد الأوروبي. تمت صياغته من قبل جوناثان برايت وأنتونيلا بيريني وآن بلوين وريجا ويس من معهد أكسفورد للإنترنت بجامعة أكسفورد

تمّ النشر في عام 2022 من قبل منظمة الأمم المتحدة للتربية والعلم والثقافة ،
7 place de Fontenoy , 75352 Paris 07 SP , France

العنوان الأصلي : *Adressing hate speech on social media: Contemporary challenges*
تمّ النشر في عام 2021 من قبل منظمة الأمم المتحدة للتربية والعلم والثقافة (اليونسكو) .

تمّ إعداد هذه الوثيقة بدعم مالي من الاتحاد الأوروبي. محتوياتها هي مسؤولية المعدّين وحدهم ولا تعكس بالضرورة وجهات نظر الاتحاد الأوروبي

التصميم الجرافيكي: Dean Dorat

CI/FEJ/2021/DP/01