محدث: مارس 2022

تُساعد WITNESS الأشخاص على استخدام الفيديو والتكنولوجيا لحماية حقوق الإنسان والدفاع عنها (witness.org). يُمكنك التعرف على المزيد عن عملنا ضد التزييف العميق والتحضير بشكل أفضل عبر الرابط: wit.to/Synthetic-Media-Deepfakes

التزييف العميق

تُسهل تقنية التزييف العميق التلاعب والتزييف بأصوات ووجوه وأفعال الأشخاص الحقيقيين، بالإضافة إلى القدرة على المطالبة بأي فيديو أو صوت مزيف. لقد أصبح الأمر مصدر قلق بالغ للمشاهير والسياسيين، وللعديد من النساء العاديات في جميع أنحاء العالم. ونتيجة لسهولة القيام بهذه الأفعال، استجابت WITNESS للأضرار الحالية واستعدت للتهديدات المستقبلية التي تركز على الفئات السكانية الضعيفة على مستوى العالم. وسنتعرف عبر المعلومات الأساسية الآتية على:

- التقنيات: ما تقنيات التزييف العميق الرئيسية وما الذي يمكنها فعله؟
- التهديدات: ما التهديدات الرئيسية المحددة عالميًا؟
- الحلول: ما الحلول التقنية والسياسات المحتملة؟

ما هي تقنية "التزييف العميق" والوسائط الاصطناعية؟

تقنية "التزييف العميق" عبارة عن أشكال جديدة من التلاعب السمعي البصري تسمح للأشخاص بإنشاء محاكاة واقعية لوجه شخص ما أو صوته أو أفعاله. فهي تمكن الأشخاص من جعل الأمر يبدو كما لو أن شخصًا ما قال أو فعل شيئًا لم يفعله أو صُنع فعل لم يحدث أبدًا. لقد أصبح صنعها أسهل، إذ تتطلب عددًا أقل من الصور المصدر لبنائها، وتُسوَّق الأدوات اللازمة لإنشائها بشكل متزايد. حاليًا، يؤثر "التزييف العميق" بشكل غير متناسب على النساء؛ لأنها تُستخدم لإنشاء صور ومقاطع فيديو جنسية غير توافقية مع وجه شخص معين. لكن هناك مخاوف من أن يؤثر "التزييف العميق" بصورة أوسع عبر المجتمع والأعمال والسياسة وكذلك في تحقيقات حقوق الإنسان وعمليات جمع الأخبار والتحقق.

إن تقنية "التزييف العميق" هي مجرد تطور واحد داخل عائلة من التقنيات التي تدعم الذكاء الاصطناعي (AI) لتوليد الوسائط

الاصطناعية. وتتيح هذه المجموعة من الأدوات والتقنيات إنشاء تمثيلات واقعية لأشخاص يفعلون أو يقولون أشياء لم يفعلوها أبدًا، أو إنشاء واقع لأشخاص وأمور لم تكن موجودة من قبل، أو لأحداث لم تحدث أبدًا.

وتتيح تقنية الوسائط الاصطناعية حاليًا هذه الأشكال من التلاعب:

- إضافة وإزالة العناصر داخل الفيديو بسهولة أكبر.
- تعديل الظروف الخلفية في مقطع فيديو، على سبيل المثال، تغيير الطقس لجعل لقطة فيديو في الصيف تظهر كما لو تم تصويرها في الشتاء.
- حركات الوجه أو الجسم المزيفة ("تحريك الدمى"): محاكاة والتحكم في تمثيل فيديو واقعي للشفاه أو تعابير الوجه أو حركة الجسم لفرد معين (على سبيل المثال لإظهار أنه كان ثملًا).
- مزامنة الشفاه المزيفة: مطابقة مقطع صوتي مع معالجة واقعية لشفاه شخص ما لجعله يبدو كما لو قال شيئًا لم يقله مطلقًا.
- صوت مزيف: إنشاء محاكاة واقعية لصوت شخص معين.
- غيّر نوع الصوت أو اجعله يبدو وكأنه شخص آخر : عدِّل صوتًا موجودًا بـ "هيئة صوت" من جنس مختلف، أو لشخص معين.
- إنشاء صورة واقعية ولكنها مزيفة تمامًا لشخص غير موجود. ويمكن أيضًا تطبيق نفس التقنية بشكل أقل إشكالية لإنشاء الهامبرغر والقطط المزيفة وما إلى ذلك.
- إنشاء صورة لحدث أو كائن من وصف نصي.
- نقل وجه واقعي من شخص إلى آخر، وهو الشكل الأكثر شيوعًا لـ "التزييف العميق".

[شاهد العديد من الأمثلة هنا]

كيفية عمل تقنية "التزييف العميق"؟

تعتمد هذه التقنيات اعتمادًا أساسيًا وليس حصريًا على شكل من أشكال الذكاء الاصطناعي يُعرف بالتعلم العميق وعمل شبكات الخصومة التوليدية أو شبكات GAN.

ولإنشاء عنصر من محتوى الوسائط الاصطناعية، عليك أن تبدأ بجمع الصور أو مصدر الفيديو للشخص أو العنصر الذي تريد تزييفه. إذ تطور GAN التزييف باستخدام شبكتين. وتنشئ إحدى الشبكات عمليات إعادة إنشاء معقولة لصور المصدر، بينما تعمل الشبكة الثانية على اكتشاف عمليات التزوير هذه. تُعاد بيانات الكشف هذه إلى الشبكة المشاركة في إنشاء عمليات التزوير، مما يمكّنها من تحسين وإنشاء نسخة مزيفة أفضل وأفضل من المصدر، على سبيل المثال وجه الشخص الذي تحاكيه.

واعتبارًا من أوائل عام 2022، لا تزال العديد من هذه التقنيات -لا سيما إنشاء تقنية التزييف العميق لمبادلة الوجه- تتطلب قوة حسابية كبيرة، وفهمًا لكيفية ضبط النموذج الخاص بك، وغالبًا ما تتطلب CGI مرحلة ما بعد الإنتاج لتحسين النتيجة النهائية. ومن الأمثلة الجيدة على تقنية التزييف العميق المتطورة التي تتطلب كل هذه المدخلات هي مقاطع فيديو "Tom Cruise " TikTok" والتي ربما تكون قد شاهدتها!

ومع ذلك، حتى مع القيود الحالية، تخدَع البشر بالفعل وسائط المحاكاة. على سبيل المثال، أظهر البحث أن الأشخاص لا يستطيعون الكشف بشكل موثوق عن الأشكال الحالية لتعديل حركة الشفاه، التي تُستخدم لمطابقة فم شخص ما بمسار صوتي جديد. ووجدت الأبحاث الحديثة أن البشر لم يكونوا قادرين على اكتشاف الوجوه الواقعية لأشخاص لم يكونوا موجودين على الإطلاق. ولا ينبغي أن نفترض أن البشر مؤهلون بطبيعتهم لاكتشاف التلاعب بالوسائط التركيبية.

المشهد الحالي للوسائط المزيفة والاصطناعية

"التزييف العميق" الضار والوسائط التركيبية -حتى الآن- ليس منتشرًا خارج نطاق الصور الجنسية غير التوافقية. وللأسف، فإن "التزييف العميق" الجنسي غير التوافقي متاح بسهولة ويُنشأ بمشاركة المشاهير أو الممثلات الإباحية أو الأشخاص العاديين.

بالإضافة إلى ذلك:

- بدأ الناس في تحدي المحتوى الحقيقي، ورفضوه باعتباره محتوى خاضعًا لتقنية "التزييف العميق".
- في حين أن هجاء تقنية "التزييف العميق" يوفر فرصًا جديدة للتعبير الحر، إلا أنه غالبًا ما يتعامل مع الخداع.
- تُستخدَم صور "الأشخاص الذين لم يكونوا موجودين من قبل" بشكل متزايد لإخفاء الحسابات المزيفة في معلومات مضللة.

التهديدات من التزييف العميق

في ورش العمل التي قادتها WITNESS بالإضافة إلى الدورات التدريبية مع أكثر من 500 شخص على مدار السنوات الثلاث الماضية، راجعنا ناقلات التهديدات المحتملة مع مجموعة من المشاركين من المجتمع المدني، بما في ذلك وسائل الإعلام الشعبية والصحفيين المحترفين ومدققي الحقائق، بالإضافة إلى الباحثين في المعلومات المضللة والتضليل. وOSINT (تحقيق مفتوح المصدر)

المتخصصين. لقد أعطوا الأولوية للمجالات التي قد تؤدي فيها أشكال التلاعب الجديدة إلى توسيع التهديدات الحالية أو إدخال تهديدات جديدة أو تغيير التهديدات الحالية أو تعزيز التهديدات الأخرى. كما سلطوا الضوء على التحديات المتعلقة بعبارة "إنها مزيفة" كقريب بلاغي لـ "إنها أخبار مزيفة".

أعطى المشاركون في اجتماعات الخبراء في البرازيل وأفريقيا جنوب الصحراء وجنوب شرق آسيا، والاجتماعات الأخرى على مستوى العالم، الأولوية لمخاوفهم الرئيسية فيما يتعلق بكيفية تأثير الأشكال الجديدة للتلاعب بوسائل الإعلام وزيادة المعلومات الخاطئة/ المضللة على عملهم ومجتمعاتهم.

- سيتعرض الصحفيون وقادة المجتمع والنشطاء المدنيون للهجوم على سمعتهم ومصداقيتهم، بناءً على الأشكال الحالية من المضايقات والعنف عبر الإنترنت التي تستهدف في الغالب النساء والأقليات. تم بالفعل شن عدد من الهجمات باستخدام مقاطع فيديو معدلة على الصحفيات، كما في حالة الصحفية الهندية البارزة رنا أيوب .
- ستواجه الشخصيات العامة الصور الجنسية غير الحسية والعنف القائم على النوع الاجتماعي بالإضافة إلى الاستخدامات الأخرى لما يسمى بضربات الشبهات ذات المصداقية. وقد يكون السياسيون المحليون معرضين للخطر بشكل خاص؛ لأن لديهم صورًا وفيرة ولكن لديهم القليل من الهيكل المؤسسي المحيط بهم، والذي يتعين على السياسيين على المستوى الوطني المساعدة في الدفاع عنه ضد هجوم إعلامي اصطناعي.
- تقويض احتمالات استخدام الفيديو كدليل على انتهاكات وجرائم حقوق الإنسان، مما يعوق المساءلة والعدالة.
- الصحفيون الذين يعانون بالفعل من زيادة التحميل ونقص الموارد لن يتمتعوا بقدرات التحليل الجنائي الإعلامي لقدرات الصحفيين ومدققي الحقائق للتحقق من الحقائق المزيفة.
- سيتم الضغط على منظمات حقوق الإنسان وجمع الأخبار والتحقق لإثبات صحة شيء ما، وكذلك لإثبات أن شيئًا ما لم يُزوَّر. ستتاح لمن هم في السلطة الفرصة لاستخدام الإنكار المعقول للمحتوى من خلال التصريح بأنه مزيف بشكل عميق.
- نظرًا لأن التزييف العميق أصبح أكثر شيوعًا وأسهل في الحجم، فسوف يساهمون في استراتيجيات "خرطوم الحريق من الباطل" التي تغمر وكالات التحقق من صحة وسائل الإعلام والتحقق من الحقائق بالمحتوى الذي يتعين عليهم التحقق منه أو فضحه. هذا يمكن أن يفرط في تشتيت انتباههم.
- ستتقاطع تقنية التزييف العميق مع الأنماط الحالية من "الحرائق الرقمية السريعة"، إذ تُشارَك الصور الخاطئة بسرعة في WhatsApp وTelegram وFacebook Messenger وتطبيقات المراسلة الأخرى.
- ستكون مؤتمرات الفيديو عبر الإنترنت عرضة للتلاعب.
- في جميع السياقات، لاحظ الأشخاص الذين استشرناهم أهمية مشاهدة التزييف العميق في سياق الأساليب الحالية للتحقق من الحقائق والتحقق منها. ستُدمَج تقنية التزييف العميق والوسائط التركيبية في حملات المؤامرة والتضليل الحالية، بالاعتماد على التكتيكات المتطورة (والردود) في تلك المنطقة.

بحثت WITNESS عن التزييف العميق والسخرية، بما في ذلك تقرير حديث Just Joking! حدد استخدامًا متزايدًا لميزة التزييف العميق للنقد الاجتماعي والسياسي القوي. لقد أظهر كيف أن تقنية التزييف العميق الساخرة ذات الصور الواقعية تلائم:

- النقد الاجتماعي: محاكاة ساخرة وسخرية لانتقاد القوة التي تحدد المشكلات الاجتماعية والسياسية والتي يعتبرها الجمهور ساخرة.
- سوء الاستخدام المتعمد : الادعاءات بأن هناك شيئًا ما ساخرًا عندما تكون معلومات مضللة وإلقاء اللوم على الجمهور لفشلهم في "الحصول على النكتة".
- سوء الاستخدام العرضي: عندما يُفقد السياق، ويُشارَك على أنه معلومات خاطئة، ويُحدَّد على أنه حقيقي.

ما الحلول المتاحة؟

هناك قدر كبير من العمل جار للتحضير بشكل أفضل للتزييف العميق. وتشعر WITNESS بالقلق بشكل عام من أن هذا العمل على "الحلول" لا يشمل بشكل كاف أصوات واحتياجات الأشخاص المتضررين من المشاكل الحالية للتلاعب بوسائل الإعلام وعنف الدولة والعنف القائم على النوع الاجتماعي والمعلومات المضللة/ التضليل في جنوب الكرة الأرضية وفي المجتمعات المهمشة في العالم. جلوبال نورث.

هل يمكننا تعليم الناس اكتشاف التزييف؟

ليس من الجيد تعليم الناس أنه يمكنهم اكتشاف التزييف العميق أو التلاعب بالوسائط الاصطناعية الأخرى. على الرغم من وجود بعض النصائح التي تساعد في اكتشافها الآن - على سبيل المثال، مواطن الخلل المرئية - فهذه مجرد أخطاء حالية في عملية التزوير وستختفي بمرور الوقت. في حال كنت ترغب في اختبار قدرتك على الرغم من ذلك، انتقل إلى: https://detectfakes.media.mit.edu/

ستعمل المنصات مثل Facebook والشركات المستقلة على تطوير أدوات يمكنها القيام ببعض الاكتشافات، لكنها ستوفر أدلة فقط ولن تكون متاحة على نطاق واسع في المستقبل القريب. من المهم أن يركز الأشخاص أيضًا على فهم التزييف العميق ضمن إطار أوسع لمحو الأمية الإعلامية ، مثل نهج SHEEP الخاص بمنظمة First Draft أو إطار عمل SIFT .

يقترح**SHEEP** (اختصار باللغة الإنجليزية) أنه لتجنب التعرض للخداع من خلال المعلومات الخاطئة عبر الإنترنت، ينبغي عليك "التفكير في **SHEEP** قبل مشاركتها."

إذ يشير حرف **(S)** إلى كلمة (SOURCE): بمعنى تحقق من صفحة حول موقع الويب أو الحساب، وانظر إلى أي معلومات عن الحساب وابحث عن الأسماء وأسماء المستخدمين.

ويشير حرف **(H)** إلى كلمة (HISTORY): بمعنى هل هذا المصدر لديه أجندة؟ اكتشف الموضوعات التي تتناولها بانتظام أو في حال كانت تروج لمنظور واحد فقط.

ويشير حرف (E) [الأول] إلى كلمة (EVIDENCE): بمعنى استكشف تفاصيل مطالبة أو ميم واكتشف ما إذا كانت مدعومة بأدلة موثوقة من مكان آخر.

ويشير حرف **(E)** إلى كلمة (EMOTION): بمعنى هل يعتمد المصدر على العاطفة لتوضيح نقطة؟ تحقق من الإحساس، واللغة التحريضية والمثيرة للانقسام.

ويشير حرف **(P)** إلى كلمة (PICTURES): بمعنى الصور ترسم ألف كلمة، حدد الرسالة التي تصورها الصورة وما إذا كان المصدر يستخدم الصور لجذب الانتباه.

يوفر SIFT نموذجًا آخر ذي صلة لتحليل الفطرة السليمة للمعلومات المشبوهة:

كيف نعتمد على القدرة الصحفية القائمة والتنسيق؟

يحتاج الصحفيون والمحققون في مجال حقوق الإنسان إلى تطوير فهم أفضل لكيفية اكتشاف التزييف العميق باستخدام الممارسات الحالية لـ OSINT ودمجها مع أدوات التحاليل الجنائية الإعلامية الجديدة التي يجري تطويرها. تعرف على المزيد في تقرير WITNESS حول الاحتياجات والعمل الأخير من الشراكة حول الذكاء الاصطناعي.

هل توجد أدوات للكشف؟ (ومن لديه حق الوصول؟)

تعمل معظم المنصات الرئيسية والعديد من الشركات الناشئة على تطوير أدوات للكشف عن "التزييف العميق".

بدأ إطلاق بعض الأدوات. مثال من Sensity.AI https://platform.sensity.ai/deepfake-detection

ومع ذلك، نوصي بالتعامل مع هذه بحذر شديد. لم تتوصل المسابقات الواسعة الأخيرة لاكتشاف التزييف العميق إلى نماذج كانت فعالة بدرجة كافية في التقنيات المعروفة أو قابلة للتطبيق بشكل كاف على التقنيات الجديدة وستكون معظم أجهزة الكشف المتاحة للجمهور أقل فعالية من الأنظمة المغلقة. تميل أدوات الكشف إلى أن تكون أقل موثوقية في حال كنت لا تعرف التقنية المستخدمة لإنشاء الوسائط التركيبية، وأقل موثوقية على الوسائط منخفضة الدقة أو المضغوطة التي نراها على الإنترنت. تُظهر تجربة حديثة لمشتبه به في وجود تقنية التزييف العميق في ميانمار تحديات الاعتماد على أجهزة الكشف المتاحة للجمهور دون مصاحبة الخبرة.

حتى مع تطوير أدوات قوية فلن تُتاح على نطاق واسع، لا سيما خارج المنصات والشركات الإعلامية المحددة. ومن المحتمل أن يتم استبعاد وسائل الإعلام ومنظمات المجتمع المدني في جنوب الكرة الأرضية، ومن المهم الدفاع عن الآليات التي تمكنهم من الوصول بشكل أكبر إلى مرافق الكشف. تناقش WITNESS من أجل زيادة المساواة في الوصول إلى أدوات الكشف، والاستثمار في مهارات وقدرات المجتمع المدني العالمي والصحافة ، وتطوير آليات التصعيد لتقديم تحليل عن التزييف العميق المشتبه به.

هل توجد أدوات للمصادقة؟ (ومن المستبعد؟)

هناك حركة متزايدة لتطوير أدوات لتتبع مصدر مقاطع الفيديو والصور بشكل أفضل - بدءًا من اللحظة التي يتم فيها تصويرها على الهواتف الذكية، إلى وقت تحريرها ثم مشاركتها أو توزيعها على وسائل التواصل الاجتماعي. يمكن أن تعرض لك "البنية التحتية للمصدر والأصالة" بعد ذلك معلومات عن مكان الصورة أو مقطع الفيديو وما إذا كان/ كيف تم تغيير صورة أو مقطع فيديو. هذا مناسب لكل من "التزييف الضحل" مثل مقاطع الفيديو ذات السياق الخاطئ أو مقاطع الفيديو المعدلة وكذلك التزييف العميق. يمكنك بعد ذلك استخدام هذه المعلومات لمساعدتك في اتخاذ قرارات بشأن الوثوق بالمحتوى. أحد الأمثلة على مبادرة في هذا المجال هو مبادرة أصالة المحتوى بقيادة Adobe، والتحالف الذي أُطلق مؤخرًا من أجل إنشاء المحتوى والأصالة (C2PA).

ومع ذلك، هناك خطر من أن الأدوات التي طُوِّرت للمساعدة بشكل أفضل في تتبع أصول مقاطع الفيديو، وإظهار كيفية التلاعب بها قد تؤدي أيضًا إلى مخاطر المراقبة والاستبعاد للأشخاص الذين لا يرغبون في إضافة بيانات ومعلومات إضافية إلى صورهم ومقاطع الفيديو، أو لا يمكنهم أن ينسبوا الصور إلى أنفسهم خوفًا مما ستفعله الحكومات والشركات بهذه المعلومات. وقادت WITNESS تقييم الأضرار وسوء الاستخدام لمواصفات C2PA من أجل تحديد هذه وغيرها من الأضرار المحتملة، ووضع استراتيجيات لتجنبها والتخفيف من حدتها. سيستمر إساءة استخدام "البنية التحتية للمصدر والأصالة"، وستجد الجهات الخبيثة ثغرات، لذا فإن الخطوة الرئيسية للمضي قدمًا هي تعزيز إطار حقوق الإنسان ، مع وجود حواجز حماية ضد الضرر، وآليات للانتصاف، وفرص لتمكين الأصوات الناقدة.

ما دور المنصات؟

إن منصات وسائل التواصل الاجتماعي مثل Facebook وTwitter لديها سياسات بشأن التزييف العميق وكيفية التعامل معها، وكذلك كيفية التعامل مع الوسائط التي تم التلاعب بها على نطاق أوسع.

تناقش WITNESS هنا سياسة Facebook وسياسةTwitter هنا

تشمل العناصر الرئيسية لهذه السياسات ما يأتي:

- هل يغطون "التزييف العميق" فقط أو الأشكال الأخرى من الوسائط التي تم التلاعب بها (على سبيل المثال، الفيديو البطيء، أو الفيديو الذي أسيء سياقه؟)
- كيف يعرّفون الضرر الناجم عن الفيديو؟
- هل النية من وراء المشاركة مهمة؟
- هل يحذفون مقطع فيديو مسيء؟ تسميه؟ توفير سياق للتلاعب؟ جعله أقل ظهورًا على موقعهم أو مشاركته بسهولة؟
- هل تنطبق على الشخصيات العامة؟

فيسبوك (ميتا)

سياسة **Facebook** محددة بشأن "التزييف العميق" بدلًا من الأشكال الأخرى للتلاعب بالفيديو أو الصور.

سيزيل Facebook الوسائط التي تم التلاعب بها عندما:

- يتم التحرير أو التصنيع -بما يتجاوز التعديلات من أجل الوضوح أو الجودة- بطرق لا تظهر للشخص العادي ومن المحتمل أن تضلل شخصًا ما ليعتقد أن أحد الأشخاص في الفيديو قال كلمات لم يقلها في الواقع.
- يتم الإنتاج باستخدام الذكاء الاصطناعي أو التعلم الآلي عبر دمج المحتوى أو استبداله أو تركيبه على مقطع فيديو، مما يجعله يبدو أصليًا.

لا تمتد هذه السياسة إلى المحتوى الساخر أو السخرية، أو الفيديو الذي عُدِّل فقط لحذف أو تغيير ترتيب الكلمات. يمكن الإشارة إلى/ أو التقاط مقاطع فيديو أخرى تلاعبت بها الجهات الخارجية المدققة للحقائق. وكانت هناك أمثلة حيث أخطأ فيسبوك في اعتبار السخرية المزيفة على أنها معلومات مضللة. أحد الأمثلة على ذلك حدث في الكاميرون عندما شارك أكاديمي وناشط محلي مقطع فيديو ملفقًا يوضح للسفير الفرنسي يخبر الكاميرونيين أنهم لم يحصلوا أبدًا على الاستقلال من الاستغلال الاستعماري لفرنسا. وصف مدققو الحقائق التابعون لجهات خارجية على Facebook في هيئة الإذاعة الفرنسية France 24 الفيديو بأنه خاطئ جزئيًا، مما أدى إلى إبطال القوة البلاغية للنقد.

ستتم إزالة الصوت أو الصور أو مقاطع الفيديو، سواء أكانت مزيفة أم لا، من Facebook في حال انتهكت أيًّا من معايير المجتمع الأخرى لدينا، بما في ذلك تلك التي تتحكم في "العري والعنف التصويري وقمع الناخبين وخطاب الكراهية ".

تويتر

سياسة تويتر متاحة هنا .

تشير سياسة تويتر إلى أنه "لا يجوز لك مشاركة الوسائط الاصطناعية أو التي تم التلاعب بها بشكل مخادع (ليس فقط التزييف العميق ولكن أيضًا أشكال التلاعب الأخرى) التي من المحتمل أن تسبب ضررًا. بالإضافة إلى ذلك، قد نصنف التغريدات التي تحتوي على وسائط تركيبية وتم التلاعب بها لمساعدة الأشخاص على فهم أصالتها وتوفير سياق إضافي".

يركز Twitter على ثلاث أسئلة رئيسة تحدد ما إذا كان بإمكانهم تصنيف المحتوى أو إزالته:

1. هل المحتوى اصطناعي أم تم التلاعب به؟
2. هل المحتوى تم مشاركته باستخدام طرق خادعة؟
3. هل من المحتمل أن يؤثر المحتوى على السلامة العامة أو يتسبب في ضرر جسيم؟

	هل من المحتمل أن يؤثر المحتوى على السلامة العامة أو يتسبب في ضرر جسيم؟	هل المحتوى تم مشاركته باستخدام طرق خادعة؟	هل المحتوى معدّل بشكل كبير ومخادع أو ملفق؟
قد يتم تصنيف المحتوى.	X	X	✓
قد يتم تصنيف المحتوى.	X	✓	X
من المحتمل أن يتم تصنيف المحتوى أو قد تتم إزالته. *	✓	X	✓
من المرجح أن يتم تصنيف المحتوى.	X	✓	✓
من المرجح أن تتم إزالة المحتوى.	✓	✓	✓

تيك توك

يحظر تطبيق TikTok عمليات التزوير الرقمية (الوسائط الاصطناعية أو الوسائط المُتلاعب بها) التي تضلل المستخدمين من خلال تشويه حقيقة الأحداث وإلحاق الضرر بموضوع الفيديو أو الأشخاص الآخرين أو المجتمع.

إجراءاتنا

ينبغي أن تكون المنصات استباقية في الإشارات، وتخفيض الترتيب - وفي أسوأ الحالات، إزالة - التزييف العميق الضار لأن المستخدمين لديهم خبرة محدودة في هذا النوع من التلاعب غير المرئي وغير المسموع بالأذن، ولأن الصحفيين ليس لديهم الأدوات الجاهزة لاكتشافها بسرعة أو بفعالية. لكن معالجة "التزييف العميق" لا تزيل المسؤولية عن معالجة الأشكال الأخرى من التلاعب بالفيديو «المزور» بنشاط مثل تسمية مقطع فيديو حقيقي بشكل خاطئ أو تحرير بسيط لمقطع فيديو حقيقي.

بعض الأسئلة المستمرة المتعلقة بالسياسات:

- كيف سيضمن كل من Facebook وTwitter وTikTok وغيرهم الكشف الدقيق عن "التزييف العميق"؟
- كيف ستصدر المنصات حكمًا عندما يكون التعديل خبيثًا، أو ما إذا كان هناك أمر ما مثل محاكاة ساخرة، أو بدلاً من ذلك يتنكر في صورة هجاء أو محاكاة ساخرة؟
- كيف سيتم توصيل ذلك إلى المستهلكين المتشككين؟
- كيف سيتأكدون من أن أي قرارات يتخذونها تخضع للشفافية والاستئناف لأنهم سيرتكبون أخطاء؟

ما هو دور المشرعون؟

بدأت الحكومات للتو في التشريع حول تقنية التزييف العميق. تتعلق هذه القوانين بكل من الصور الجنسية غير التوافقية، وكذلك استخدام الخداع/ التضليل.

في الولايات المتحدة، تم اقتراح عدد من القوانين على مستوى الولاية والمستوى الفيدرالي، وفي الاتحاد الأوروبي يؤيد قانون الذكاء الاصطناعي وضع العلامات على الوسائط التركيبية لحماية المستهلك.

في منطقة آسيا والمحيط الهادئ، هناك مثالان على ذلك هما القوانين في جمهورية الصين الشعبية، التي تحظر التزييف العميق وغيرها من "الأخبار المزيفة"، والتشريعات المقترحة مؤخرًا في الفلبين. أحد التحذيرات بشأن هذه القوانين هو عندما تضع تعريفًا واسعًا جدًّا للتزوير السمعي البصري وتتضمن أشكالًا مهمة من حرية التعبير مثل السخرية، أو تمنح الحكومات حرية التصرف والسلطة لتقرير ما هو "مزيف".