Cách YouTube bảo vệ cộng đồng khỏi những lời nói thù địch và hành vi quấy rối

Tại YouTube, chúng tôi muốn trao cho mọi người tiếng nói. Những lời nói thù địch và hành vi quấy rối ngăn mọi người chia sẻ câu chuyện và trải nghiệm của bản thân với thế giới. Vì thế, chúng tôi đã xây dựng các chính sách ngăn chặn nội dung như vậy trên YouTube. Tương tự như biện pháp đối với những nội dung gây hại khác, chúng tôi nhanh chóng gỡ bỏ nội dung vi phạm chính sách về lời nói thù địch và hành vi quấy rối.

Các chính sách của YouTube bảo vệ ai?

Chúng tôi phân biệt lời nói thù địch và hành vi quấy rối theo mục đích của nội dung sai trái.

Chính sách về lời nói thù địch bảo vệ các nhóm người

Chính sách của chúng tôi về lời nói thù địch hạn chế bất kỳ nội dung nào kích động bạo lực hoặc khuyến khích thái độ hận thù đối với các nhóm người dựa trên các đặc điểm được bảo vệ.

Chính sách về hành vi quấy rối bảo vệ các cá nhân

Chính sách về hành vi quấy rối hạn chế những lời đe dọa nguy hại hoặc lời lăng mạ nhắm đến các cá nhân có thể nhận dạng được dựa trên các đặc điểm riêng biệt, bao gồm cả các đặc điểm được bảo vệ hoặc đặc điểm cơ thể.

Các đặc điểm được bảo vệ

Định nghĩa của chúng tôi về những nhóm người này dựa trên những luật hiện hành cũng như thông qua việc tham khảo ý kiến của chuyên gia trong các lĩnh vực như quyền công dân và lời nói thù địch.

- ✓ Chủng tộc
- ✓ Sắc tộc
- ✓ Quốc tịch
- ✓ Tôn giáo
- ✓ Tình trạng khuyết tật
- ✓ Giới tính
- ✓ Xu hướng tính dục
- ✓ Tình trạng nhập cư
- ✓ Độ tuổi
- ✓ Địa vị xã hội
- ✓ Tình trạng cựu chiến binh
- ✓ Tình trạng là nạn nhân của một sự kiện bạo lực lớn (hoặc người thân của họ)

Các chính sách của YouTube bảo vệ cộng đồng khỏi những mối nguy nào?

Dưới đây là các ví dụ về nội dung không được phép đăng trên YouTube (nội dung tham khảo từ chính sách về lời nói thù địch và hành vi quấy rối trên mạng).

Lời nói thù địch

- ✕ Thuyết âm mưu dùng để biện minh cho hành vi bạo lực trong thế giới thực
- ✕ Hành vi hạ thấp nhân phẩm
- ✕ Hành vi kích động bạo lực
- ✕ Hạ thấp địa vị
- ✕ Thuyết ưu thế
- ✕ Hành vi đe dọa
- ✕ Phủ nhận sự kiện bạo lực

Hành vi quấy rối

- ✕ Tiết lộ thông tin cá nhân (Doxxing)
- ✕ Cảnh mô phỏng hành động bạo lực
- ✕ Hành vi đeo bám
- ✕ Hành vi đe dọa
- ✕ Hành vi tình dục cưỡng ép
- ✕ Hành vi quấy rối nạn nhân của một vụ bạo lực
- ✕ Ý muốn chết
- ✕ Hành vi hạ thấp nhân phẩm
- ✕ Lời lăng mạ quá khích
- ✕ Lời lăng mạ nhắm vào một cá nhân
- ✕ Nội dung nhắm vào những cá nhân dễ bị tổn thương

Chúng tôi chuyển đổi như thế nào để xử lý các hình thức mới của lời nói thù địch và hành vi quấy rối?

Trong mọi ngôn ngữ, những câu nói thể hiện lời nói thù địch và hành vi quấy rối thay đổi liên tục. Chúng tôi có các bước để đảm bảo chính sách của mình không ngừng biến đổi để đối phó với các hình thức, ngôn từ và mục tiêu mới của nạn phân biệt.

Bộ phận xử lý dữ liệu là nhóm nội bộ quản lý các nguồn tin tức, mạng xã hội và nội dung bị người dùng báo vi phạm

Chúng tôi hợp tác với các chuyên gia bên ngoài thuộc các lĩnh vực như chủ nghĩa bạo lực cực đoan, thuyết ưu thế, quyền công dân, hành vi bắt nạt qua mạng và quyền tự do ngôn luận

Chúng tôi liên tục tham khảo ý kiến của các nhà sáng tạo về trải nghiệm của họ để phát triển và cập nhật chính sách

YouTube thực thi các chính sách như thế nào?

Cộng đồng và hệ thống máy học giúp chúng tôi phát hiện lời nói căm thù và hành vi quấy rối

Vì mỗi phút có hơn 500 giờ video được tải lên YouTube, nên chúng tôi để người xem và các nhà sáng tạo có quyền áp dụng biện pháp xử lý, đồng thời đầu tư vào các hệ thống máy học nhằm phát hiện ra nội dung có vấn đề ở quy mô lớn.

Chương trình Người gắn cờ đáng tin cậy cho các đối tác chính phủ và tổ chức phi chính phủ có mức độ chính xác cao trong việc báo cáo lời nói căm thù và hành vi quấy rối

Nhà sáng tạo có thể chặn một số từ ngữ hoặc các cá nhân trong phần bình luận bằng công cụ kiểm duyệt bình luận

Người xem có thể gắn cờ các video và bình luận

Công nghệ máy học giúp chúng tôi phát hiện nội dung vi phạm ở quy mô lớn. Hệ thống của chúng tôi được huấn luyện để phát hiện nội dung gây hại trên YouTube 24/7. Nhờ đó, chúng tôi có thể xem xét hàng trăm nghìn video chỉ trong một thời gian rất ngắn so với một người thường.

Nhân viên đánh giá nội dung đánh giá nội dung bị gắn cờ

Nội dung bị gắn cờ không bị hệ thống tự động gỡ bỏ. Nhóm chuyên gia đánh giá của chúng tôi trên toàn thế giới sẽ quyết định nội dung cần gỡ bỏ dựa trên chính sách về lời nói căm thù và hành vi quấy rối. Quyết định của họ được dùng để huấn luyện và cải thiện hệ thống máy học.

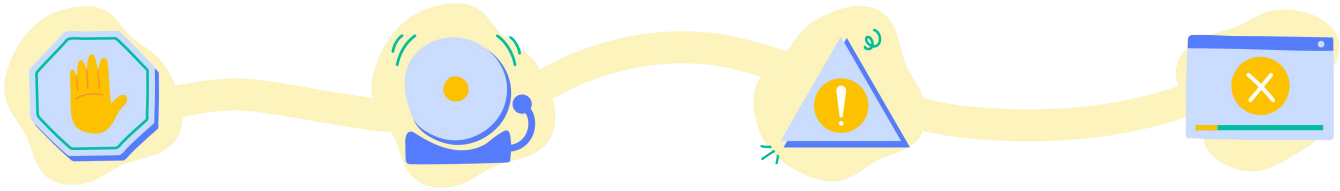
Thông thạo hàng trăm thứ tiếng trên toàn thế giới và kỹ lưỡng trong việc đánh giá nội dung bị gắn cờ

Huấn luyện và đánh giá để hiểu rõ các sắc thái nghĩa của bối cảnh hoặc ngôn ngữ

Thực thi các nguyên tắc một cách nhất quán, dù người sáng tạo có khác nhau về lai lịch, quan điểm chính trị, địa vị hay liên minh

Gỡ bỏ nội dung vi phạm chính sách

Hình thức xử phạt những nhà sáng tạo đăng tải nội dung thù địch và quấy rối



1. **Thông báo cho nhà sáng tạo** biết những biện pháp chúng tôi áp dụng và lý do áp dụng
2. **Nhắc nhở** đối với lần đầu tiên vi phạm Nguyên tắc cộng đồng
3. **Cảnh cáo** đối với các lần vi phạm tiếp theo, đồng thời không cho phép người sáng tạo đăng nội dung trong ít nhất 1 tuần
4. **Chấm dứt kênh** nếu kênh nhận 3 cảnh cáo trong vòng 90 ngày hoặc trong khoảng thời gian ngắn hơn nếu có hành vi vi phạm nghiêm trọng

Trường hợp ngoại lệ không theo chính sách

Chúng tôi có thể áp dụng ngoại lệ không theo chính sách đối với nội dung thú vị nhằm cung cấp kiến thức hoặc tài liệu, phục vụ khoa học hoặc nghệ thuật, gọi là nội dung "EDSA". Ví dụ: Chúng tôi có thể cho phép nội dung hài kịch hoặc trào phúng sử dụng những lời lăng mạ hay ngôn từ phản cảm khác nếu bối cảnh được thể hiện rõ ràng cho người xem.

Hạn chế việc lan truyền nội dung gần ranh giới vi phạm chính sách

Một số nội dung đến gần ranh giới vi phạm chính sách về lời nói căm thù và hành vi quấy rối, nhưng chưa vượt quá ranh giới. Chúng tôi hạn chế việc lan truyền nội dung gần ranh giới vi phạm chính sách bằng cách giới hạn việc đề xuất nội dung đó trên trang chủ của người xem và danh sách "Tiếp theo".

YouTube đã tiến triển như thế nào trong việc xóa nội dung có lời nói thù địch và hành vi quấy rối?

Nhờ áp dụng công nghệ, kiến thức chuyên môn và chính sách thích hợp, chúng tôi có thể gỡ bỏ các video, kênh và bình luận có chứa lời nói thù địch và hành vi quấy rối nhanh hơn bao giờ hết.

Lời nói thù địch*

Hơn 88 nghìn video bị gỡ bỏ

Hơn 43 triệu bình luận bị gỡ bỏ

Hành vi quấy rối*

Hơn 322 nghìn video bị gỡ bỏ

Hơn 135 triệu bình luận bị gỡ bỏ

70%

video có dưới 10 lượt xem bị gỡ bỏ do vi phạm chính sách*

youtube.com/howyoutubeworks

*Từ tháng 10/2020 đến tháng 12/2020