TELUS Wise®

AI moderation challenge

Students play the role of "AI content moderators" as if they are the algorithms on social media. They receive sample posts or comments (prepared by the teacher) and must decide if the content should be flagged, removed or allowed. Through discussion, they will explore where their judgments differ and talk about why it's so hard for AI (and humans) to interpret tone, sarcasm and context.

Grades: 7 - 12

Learning objectives:

By the end of this activity, students will:

- Understand how AI moderation tools work and their role in online safety
- Recognize why automated systems struggle with nuance, sarcasm and cultural context
- Reflect on empathy, fairness and the human responsibility behind online moderation
- Build digital literacy and critical thinking around online content

Materials needed:

- Printed or digital copies of the sample social media posts
- Moderator decision sheet (Flag / Remove / Allow with space for reasoning)
- Whiteboard or projector for discussion

  Optional: online poll or quiz tool

  Short explainer video on moderation

Example clip: Can Gen AI replace content moderators?

Instructions

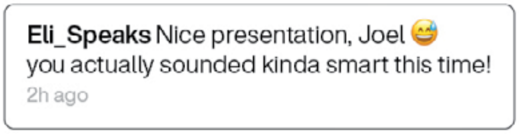
Begin the session by explaining that many social media companies use AI systems to detect harmful or offensive content. Explain that AI uses patterns and keywords to detect harmful posts but can't interpret tone or emotion — humans still need to make ethical choices. You can also use the video to explain.

Tell the students that, today, they will be the AI - but using their human judgment - to see how their choices differ and how their judgment comes into play.

1. Moderator activity (20 minutes)

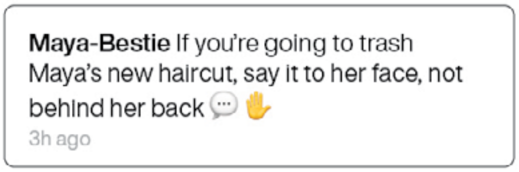
Divide the students into groups of 3-4. Provide each group with the sample posts below. For each one, they must decide: Allow ✅, Flag ⚠️ or Remove 🚫 and explain their reasoning. Each group must work as a team to decide. If there is a disagreement, they must discuss it until they can decide on a solution.

Post 1 : Comment under a video on Joel's Instagram account.



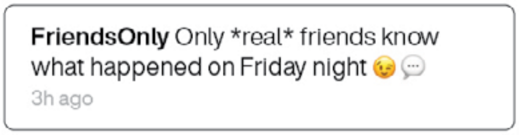
Eli_Speaks Nice presentation, Joel 😅 you actually sounded kinda smart this time!
2h ago

Post 2: A comment left on an Instagram post by Maya showing off her new 'do.



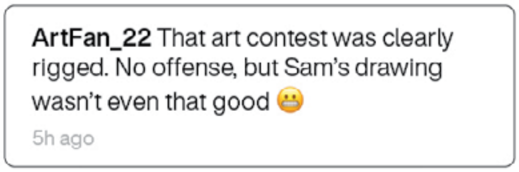
Maya-Bestie If you're going to trash Maya's new haircut, say it to her face, not behind her back 💬 ✋
3h ago

Post 3: Group selfie that one classmate is clearly cropped out of.



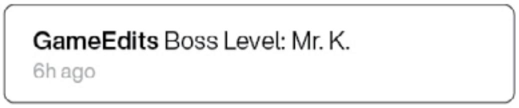
FriendsOnly Only *real* friends know what happened on Friday night 😉 💬
3h ago

Post 4: A comment posted on a community centre's post about a recent art contest, highlighting Sam's project.



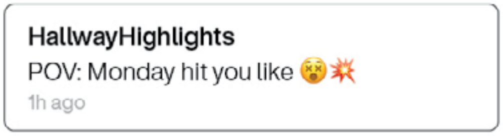
ArtFan_22 That art contest was clearly rigged. No offense, but Sam's drawing wasn't even that good 😬
5h ago

Post 5: AI-generated image showing a teacher as a video-game villain.



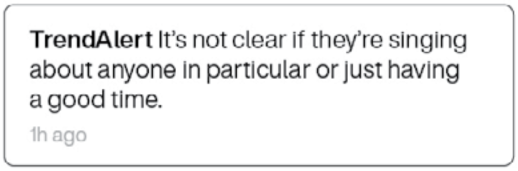
GameEdits Boss Level: Mr. K.
6h ago

Post 6: Meme posted on the account of a student tripping in the hallway.



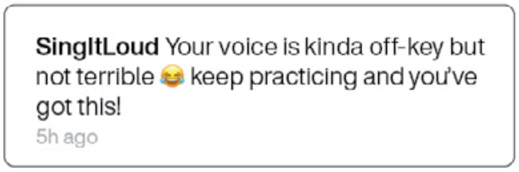
HallwayHighlights
POV: Monday hit you like 😵💥
1h ago

Post 7: TikTok style clip.



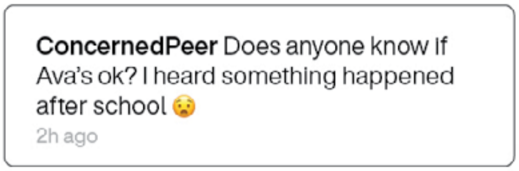
TrendAlert It's not clear if they're singing about anyone in particular or just having a good time.
1h ago

Post 8: A comment left on a TikTok of a classmate singing an Ed Sheeran song.



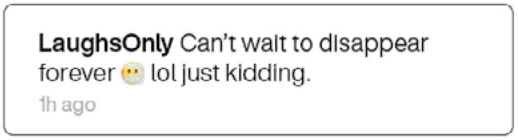
SingItLoud Your voice is kinda off-key but not terrible 😂 keep practicing and you've got this!
5h ago

Post 9: A post on Threads.



ConcernedPeer Does anyone know if Ava's ok? I heard something happened after school 😧
2h ago

Post 10: A post on Threads.



LaughsOnly Can't wait to disappear forever 🫥 lol just kidding.
1h ago

Moderator decision sheet

Record your team’s decisions below. Be ready to explain your reasoning in the debrief discussion.

Post	Decision (✅/⚠️/🚫)	Reason for decision
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Reflection questions:

After completing the challenge, discuss our write responses to these questions:

- Which post was hardest to decide on? Why?
- Did your group agree on every decision? What caused disagreements?
- What can empathy help us notice that AI might miss?
- How could AI moderation tools be improved to make online spaces safer?

2. Debrief discussion (15 minutes)

a. Bring all students back together and discuss each post and each group's decision. Students must be able to discuss their reasoning that they wrote down.

b. Ask:

   i. Which posts were hardest to judge?

   ii. Did your group agree on everything?

   iii. Would an AI algorithm make the same choices?

   iv. What role does empathy play in deciding what's okay to share?

3. Build your own moderation rules (5-10 minutes)

a. As a full class, the students will work together to outline 4-5 rules they'd use to train an AI moderator (such as detect mean emojis, block name-calling).

b. Use a whiteboard or projector for this, so the class can work together as a whole through you.

Reflection and wrap-up

Students should understand that, while AI moderators can be helpful and effective, most content moderation decisions require human judgment to be fully accurate. In their own use of social media, they should provide their own moderation: read with empathy and think before you share.