GAME PURPOSE

By playing the role of both a startup social media platform policy trust and saftey team and a content moderator, participants can begin to experience some of the challenges associated with moderating user generated online content in a way that balances values such as free expression and community safety.

Game works best with **3** players, but other sized groups can still play

THE STORY

Congratulations, your brand new social media startup, Contentr has just received funding from investors! You've been following the news and are determined to avoid the same mistakes as your predecessors, so the first place you want to start is to develop your content moderation policy. There are three rounds to the game: first you'll work as a team to develop the policy that will help you shape the kind of platform you want to grow. You'll then switch roles from policy developer to content moderator, where you'll use your policy to make moderation decisions, based on real life examples. Finally, you'll see how your decisions play out, calculating your final score based on real life examples of moderation decisions, and how those decisions affect two areas: free expression and community safety.

You will begin with 500 free expression points and 500 community safety points. Free expression points are important because they provide space for your users to express themselves and community safety points are important because they ensure your users are free from potential off-platform harms. The more you are able to balance your points, the more "mass appeal" your game will have, resulting in more ad revenue.

CONTENT WARNING

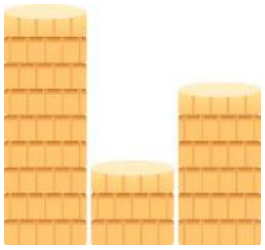
This game involves discussing descriptions of (but not viewing) sexually explicit content, harassment, hate speech, self-harm, illegal activity, misinformation, and violent content. The purpose of the game is to provide deeper understanding of platform governance which is inherently challenging, frustrating, and sometimes upsetting, and these are emotions you may feel as you play the game. You should take breaks and feel free to leave the game if needed.

Additionally, we encourage the use of John Stavropoulos' X-card strategy (link here: http://tinyurl.com/x-card-rpg). If a card is particularly uncomfortable to engage with, simply press this button:

X

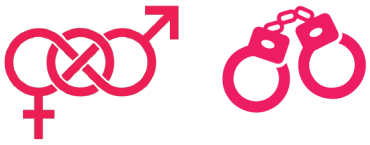
BEGIN: SETUP

The game is divided into three cycles, each focuses on different areas of controversial content. At the start of each cycle, you will receive an investment which will allow the Contentr Trust & Safety team to make decisions related to growth. Throughout the game you will experience changes to both free expression and community safety points and your budget.



CYCLES

You will notice, the cycles are grouped by category of controversial content and the cycle is displayed on the card:



Cycle 1:
Sexually Explicit Content & Illegal Activity

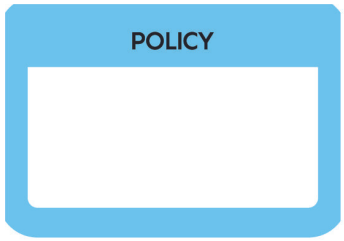
Cycle 2:
Self Harm & Graphic Content

Cycle 3:
Harassment, Hate Speech & Quality Contributions

Depending on time you may choose to only play one cycle. Each cycle lasts approx. 60 minutes. For the deepest experience, play the cycles in order, but playing a single cycle is also fun.

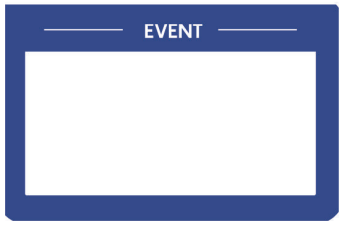
ROUNDS

Each cycle has rounds which are signaled by the border of the cards.



Round 1: Policy
Playing the role of the Trust and Safety team, write Contentr's community guidelines/platform policies

Round 2: Content
Playing the role of the Content Moderator, enact the Contentr's policies (one nuance card per cycle)



Round 3: Event
Events based on real life examples of pushback social media companies have forced over the years. Some consequences are based on decisions made in rounds 1 and 2, and other events happen regardless of the platform's efforts

PLAY YOUR FIRST CYCLE

You may be wondering, "wait, what is Contentr?"

Contentr is a start up social media company attempting to take market share from the today's leading platforms. The site allows users to connect with friends, community members and high profile figures. It includes a range of user generated content: text, photos, videos, articles. Lastly, it includes ads and a news feed.

The company's values will be shaped as you form policies.

ROUND 1: SET POLICIES

In this round, you will decide what kind of content your company wants to allow, and what it wants to prohibit.

To start, Find the description of the Trust and Safety role and review the card.

SORT

You'll all work together to decide which policies to enact, but in the interest of time, since there are three of you, when you disagree, majority rules.

Rather than starting from scratch, you'll select from a set of existing policies. (It is very common for new platforms to look at the policies of competitor social media platforms for inspiration.) You'll do this by sorting the light blue policy cards into one of two categories: Allowed or Banned.

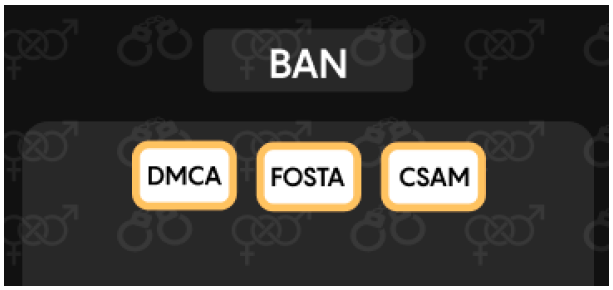
SECTION 230 & PRE CYCLE

To guide you, you have your own values as a company, but you also need to follow existing legislation. First is Section 230 of the Communications Decency Act, which provides your team with a “shield” and a “sword.” The shield means that your platform cannot be held liable for content users post with a few exceptions (pre-banned cards, discussed below). The sword means your platform can “moderate” content on the platform as long as you do not “publish” or “edit.”

But there are some exceptions, represented with pre-banned cards. As a small company you will do your best to avoid this content, as a large legal battle may bankrupt you.

- The Digital Millennium Copyright Act (DMCA) states intermediaries such as your platform have “indirect liability” if they do not make efforts to take down copyrighted material.
- The Fight Online Sex Trafficking Act (FOSTA), was passed in 2018 and outlaws content that facilitates traffickers in advertising the sale of unlawful sex acts with sex trafficking victims.
- Federal law requires that you take down and report content representing child sexual exploitation as soon as reasonably possible. Represented as a reference to Title 18 crime laws (TX VII)

Other than these three types of banned content, you have the shield and the sword - moderation decisions are up to you.

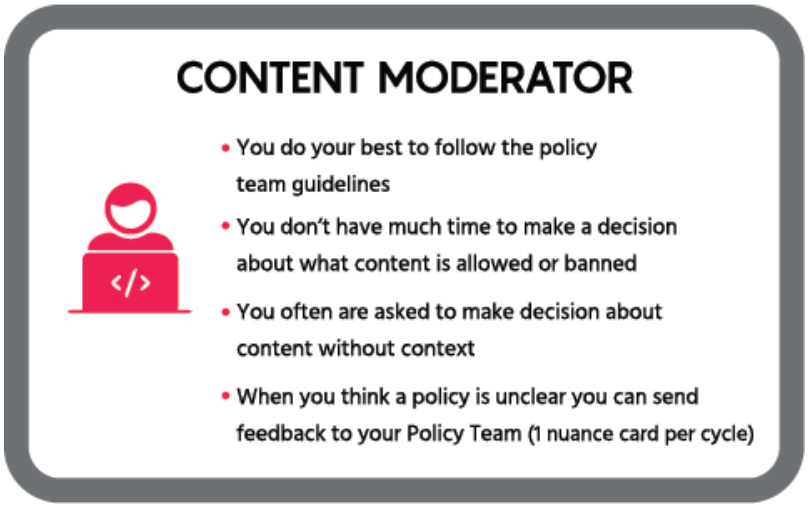


ROUND 2: MODERATING CONTENT

Excellent work! You now have a policy that you'll use to make decisions about the kind of content that's allowed on your platform. In this round you'll switch roles, from policy developer to content moderator.

Find the description of the Content Moderator role and review the card.

Once you understand the role, sort the grey cards into allow or banned based on the policy cards sorted in round 1.



NUANCE CARDS

Finally, as in real life, a content moderator can suggest policy changes to your company's trust and safety team based on their personal experiences. At the end of the cycle, you will be able to add nuance to one existing policy. For example, if a gray card is placed in the "allow" or "banned" pile based on existing policies and you want to switch it, you can edit a policy to make the switch.

ROUND 3: CONSEQUENCES

You've made it to round 3! All the hard work is done and now you get to see the outcomes of your decisions. You started with 500 free expression (FE) points, and 500 community safety (CS) points.

In this round you'll read the event cards (Dark Blue Border). The blue cards are based on examples of pushback social media companies have faced over the years. Some consequences are based on decisions made in rounds 1 and 2, and other events happen regardless of the platform's efforts.

You will notice different types of event cards, including "Society" (events that occurred in the society at large) and "algorithm" (automated decision-making trained rightly or wrongly based on decisions made during rounds 1 and 2).

Events are based on real world push back that leading social media platforms have received. To read the reference press the "more"

OPTIONS TO COMPLETE THE GAME & REFLECTION

Option: play another cycle.

If you choose to play another cycle: Stack the cards currently in your “Ban” column and place next to the pink “Ban” card. Stack the cards currently in your “Allow” column and place them next to the pink “Allow” card.

After Cycle 3...

Option: The game is complete.

Take a look at your final score (both the FE/CS points and revenue). Do you feel like you “won”? why? why not?

Takeaways from the game:

- How did your views of social media change after playing the game?
- Which decisions were easy, why?
- Which decisions were challenging, why?

BACKGROUND

Social media content moderation practices vary from company to company, are inherently opaque and span well beyond simply allowing or banning content. This game is meant to give a taste of the challenges posed by hosting a site for user generated content but should in no way be interpreted as a comprehensive overview of trust and safety practices.

The categories of content and roles were informed by:

Block, Hans, et al. (2018).The Cleaners. (film)

Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Roberts, S. T. (2019). Behind the screen: Content moderation in the shadows of social media. Yale University Press.

ACKNOWLEDGEMENTS

The game design research was sponsored by the National Science Foundation under award CNS-1452854. UMD IRB 1682807-2.

Developer of the Virtual Game: Devin Navas

Anna Lenhart, PhD Student, College of Information Studies, University of Maryland

Dr. Sarah Gilbert, Postdoctoral Associate, College of Information Studies, University of Maryland

Dr. Katie Shilton, Associate Professor, College of Information Studies, University of Maryland

Special thanks to everyone who participated in game design trials!

CARDS INSPIRED BY

(2016, May 20). 'A warning to other states': PETA wins $250,000 over Idaho 'Ag-gag' case. RT. https://www.rt.com/usa/343834-peta-paid-fees-ag-gag/

(2019, April 23). Bernie Sanders vows to round up remaining ISIS members, allows them to vote. Babylon Bee. https://babylonbee.com/news/bernie-sanders-vows-to-round-up-remaining-isis-members-allow-them-to-vote

Alba, D. (2021, March 19). How Anti-Asian Activity Online Set the Stage for Real-World Violence. New York Times. https://www.nytimes.com/2021/03/19/technology/how-anti-asian-activity-online-set-the-stage-for-real-world-violence.html

Anderson, C. A., & Carnagey, N. L. (2009). Causal effects of violent sports video games on aggression: Is it competitiveness or violent content?. Journal of experimental social psychology, 45(4), 731-739.

Asher-Schapiro, A. (2017, Nov 2). YouTube and Facebook are removing evidence of atrocities, jeopardizing cases against war criminals. The Intercept. https://theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya/

Arnold, J. (2019, May 23). Artist who turns MAGA hats into symbols of hate speech banned from Facebook. WUSA9. https://www.wusa9.com/article/news/md-artist-banned-from-facebook-for-controversial-maga-hat-art/65-1cd87d9d-2761-4ff8-a3e4-209919489ccb

Booth, R. & Weaver, M. (2015, Jun 5). 'Baby yoga' video on Facebook sparks internet censorship debate. The Guardian. https://www.theguardian.com/technology/2015/jun/05/baby-yoga-video-facebook-internet-censorship-debate

Bullen, S. (2019, Aug 26). Instagram's Graphic Self-Harm Content Ban is not Enough. PublicEar.https://medium.com/the-public-ear/instagrams-graphic-self-harm-content-ban-is-not-enough-5df3060f41cc

REFERENCES

CARDS INSPIRED BY

Brandom, R. & Newton, C. (2017, Feb 24). Twitter is locking accounts that swear at famous people. The Verge. https://www.theverge.com/2017/2/24/14719828/twitter-account-lock-ban-swearing-abuse-moderation

DANTE [@1917Dante]. (2020, Jun 6). @jack @Twitter#BLM #BlackLivesMattters #BlackLivesMatterDC #DCProtests #DC #DCProud My Twitter was blocked for almost 7 days because of the following tweet and photos. Can you please explain me how I violated the twitter rules? [Image attached] [Tweet]. Twitter. https://twitter.com/1917Dante/status/1269390501880496136

Delfino, R. A. (2020). Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. Actual Probs. Econ. & L., 105.

Farokhmanesh, M. (2018, Jun 4). YouTube is still restricting and demonetizing LGBT videos – and adding anti-LGBT ads to some. The Verge. https://www.theverge.com/2018/6/4/17424472/youtube-lgbt-demonetization-ads-algorithm

Frishberg, H. (2019, Oct 29). 'Sexual' use of eggplant and peach emojis banned on Facebook, Instagram. New York Post. https://nypost.com/2019/10/29/sexual-use-of-eggplant-and-peach-emojis-banned-on-facebook-instagram/

Galindo, Y. (2017, Oct 25). Machine Learning Detects Marketing and Sale of Opioids on Twitter. UC San Diego News Center. https://ucsdnews.ucsd.edu/pressrelease/machine_learning_detects_marketing_and_sale_of_opioids_on_twitter

Gillbert, B. (2019, Nov 6). The 10 most-viewed fake-news stories on Facebook in 2019 were just revealed in a new report. Insider. https://www.businessinsider.com/most-viewed-fake-news-stories-shared-on-facebook-2019-2019-11

CARDS INSPIRED BY

Gore, I. (2016, May 7). Illma Gore:'If anyone is going to be threatened by a small penis, it's Trump' The Guardian. https://www.theguardian.com/us-news/2016/may/07/donald-trump-penis-painting-ilma-gore

Hamilton, I.A. (2018, Aug 30). Facebook apologises to the Anne Frank Center for removing image of naked child holocaust victims. Insider. https://www.businessinsider.com/facebook-apologises-for-removing-anne-frank-center-child-holocaust-image-2018-8

Hesse, J. (2016, Feb 17). Facebook cracks down on marijuana firms with dozens of accounts shut down. The Guardian. https://www.theguardian.com/us-news/2016/feb/17/facebook-marijuana-cannabis-businesses-crackdown

Kesslen, B. (2019, Feb 4) Instagram to hide self-harm images in the wake of rising teen suicides. NBC News. https://www.nbcnews.com/news/us-news/instagram-hide-self-harm-images-wake-rising-teen-suicides-n966781

Johnson, M. (2016, Dec 14). How fake news led Dylann rook to murder nine people. The Undefeated. https://theundefeated.com/features/how-fake-news-led-to-dylann-roof-to-murder-nine-people/

Keeley, M. (2019, Aug 22). YouTube's AI Flagged Robot Battles as Animal Cruelty and Removed Them. Newsweek. https://www.newsweek.com/youtubes-ai-flagged-robot-battles-animal-cruelty-removed-them-1455806

Kelly Garrett, R (2019, Aug 16). Too many people think satirical news is real. The Conversation. https://theconversation.com/too-many-people-think-satirical-news-is-real-121666

Lorenz, T. (2017, Dec 5). Facebook is Banning Women for Calling Men 'Scum.' Daily Beast. https://www.thedailybeast.com/women-are-getting-banned-from-facebook-for-calling-men-scum

Mackey, A. (2020, Sep 17). Plaintiffs Continue Effort to Overturn FOSTA, One of the Broadest Internet Censorship Laws. EFF. https://www.eff.org/deeplinks/2020/09/plaintiffs-continue-effort-overturn-fosta-one-broadest-internet-censorship-laws

CARDS INSPIRED BY

Matney, L. (2018, May 25). Facebook has a very specific Pepe the Frog policy, report says. TechCrunch. https://techcrunch.com/2018/05/25/facebook-has-a-very-specific-pepe-the-frog-policy-report-says/

Magistretti, B. (2018, April 5). Facebook's ad policies are hurting women's heath startups. VentureBeat. https://venturebeat.com/2018/04/05/facebooks-ad-policies-are-hurting-womens-health-startups/

Melendez, S. (2020, March 3). 'I have a duty to do this': Meet the Redditors fighting 2020's fake news war. Fast Company. https://www.fastcompany.com/90466966/i-have-a-duty-to-do-this-meet-the-redditors-fighting-2020s-fake-news-war

Messent, P. (2011, Jan 5). Censoring Mark Twain's 'n-words' is unacceptable. The Guardian. https://www.theguardian.com/books/booksblog/2011/jan/05/censoring-mark-twain-n-word-unacceptable

Mpsos, N. (2017, Feb 24). Instagram bans 'nude' video of Himba woman. IOL. https://www.iol.co.za/capetimes/arts-portal/instagram-bans-nude-video-of-himba-woman-7913679

Notopoulos, K. (2017, Dec 2). How Trolls Locked my Twitter Account for 10 Days, and Welp. BuzzFeed.News.https://www.buzzfeednews.com/article/katienotopoulos/how-trolls-locked-my-twitter-account-for-10-days-and-welp

Oppel, R. (2013, Apr 7). Taping of Farm Cruelty is Becoming the Crime. New York Times. https://www.nytimes.com/2013/04/07/us/taping-of-farm-cruelty-is-becoming-the-crime.html

Oberhaus, D. (2018, Aug 29). Life on the internet is hard when your last name is 'Butts'. Vice. https://www.vice.com/en/article/9kmp9v/life-on-the-internet-is-hard-when-your-last-name-is-butts

REFERENCES

CARDS INSPIRED BY

Paul, K. (2020, May 12). Facebook to pay $52m for failing to protect moderators from 'horrors' of graphic content. The Guardian. https://www.theguardian.com/technology/2020/may/12/facebook-settlement-mental-health-moderators

Patel M, Lee AD, Clemmons NS, et al. National Update on Measles Cases and Outbreaks — United States, January 1–October 1, 2019. MMWR Morb Mortal Wkly Rep 2019;68:893–896. DOI: http://dx.doi.org/10.15585/mmwr.mm6840e2

Parkinson, H.J. (2015, Nov 3). A surprisingly difficult question for facebook: do I have boobs now? The Guardian. https://www.theguardian.com/technology/2015/nov/03/facebook-instagram-do-i-have-boobs-now

Peterson, A. (2016, Jul 7). Why the Philando Castile police-shooting video disappeared from Facebook – then came back. Washington Post. https://www.washingtonpost.com/news/the-switch/wp/2016/07/07/why-facebook-took-down-the-philando-castile-shooting-video-then-put-it-back-up/

R/AMA - comment by U/Appls on "I did content moderation for Facebook for almost a year. Ama". reddit. (n.d.). Retrieved November 6, 2021, from https://www.reddit.com/r/AMA/comments/90f3dv/i_did_content_moderation_for_facebook_for_almost/e2pyq3e

Richards, K. (2018, Jan 26). Save Your Apologies: Facebook Deletes Black Woman's Post About White Women. BLAVITY: NEWS. https://blavity.com/save-your-apologies-facebook-deletes-black-womans-post-about-white-women?category1=feminism&subCat=news&category2=news

Samakow, J. (2017, Dec 6). 'Stop Censoring Motherhood' Movement Takes Aim At Facebook and Instagram. Huffpost. https://www.huffpost.com/entry/facebook-instagram-censoring-motherhood_n_5578300

Sex School [@SexSchoolHub]. (2019, Jan 26). So we tried to post the photo of this potato on Instagram. AND IT GOT CENSORED Hey @instagram Since when potatoes are not allowed on your platform?? #Instagram #censorship #potatoban [Image attached] [Tweet]. Twitter. https://twitter.com/SexSchoolHub/status/1089176609792376834

REFERENCES

CARDS INSPIRED BY

Szalavitz, M. (2019, July 2). Facebook is censoring posts that could save opioid users' lives. Vice. https://www.vice.com/en/article/qv75ap/facebook-is-censoring-harm-reduction-posts-that-could-save-opioid-users-lives

Tiffany, K. (2019, Jun 19). The Hired guns of Instagram: Companies can't advertise on social media—so they have female influencers do it for them. Vox. https://www.vox.com/features/2019/6/19/18644129/instagram-gun-influencers-second-amendment-tactical-community

Wegmann, P. (2018, Feb 27). Google tried censoring 'gun' shopping searches. It backfired. Washington Examiner. https://www.washingtonexaminer.com/google-tried-censoring-gun-shopping-searches-it-backfired

When you find out your daughter likes black guys Whatca Gonna do brother?: Wrestling meme on Me.me. me.me. (n.d.). Retrieved November 6, 2021, from https://me.me/i/when-you-find-out-your-daughter-likes-black-guys-whatca-15048559

Wong, J.C. (2017, May 19). Facebook blocks Pulitzer-winning reporter over Malta government expose. The Guardian. https://www.theguardian.com/world/2017/may/19/facebook-blocks-malta-journalist-joseph-muscat-panama-papers

Yu, Y. (2021, March 26). Social media chiefs grilled in US Congress over anti-Asian content. Nikkei Asia. https://asia.nikkei.com/Business/Technology/Social-media-chiefs-grilled-in-US-Congress-over-anti-Asian-content