Appendix B:
Facilitation Guide: Content Moderation Policy by Design Game

Game Purpose: By playing the role of both a startup social media platform policy 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 free expression values and community safety.

Facilitation rules:

- Remind participants of roles if they begin speaking too long about personal experience
- Ask questions when someone makes a statement without explanation ("say more?" or "what do you mean by that?")
- Re-read the policy cards relevant to the borderline content when the group is stuck
- 2-3 minutes or as fruitful discussion occurs – call a vote

Orientation/Warning:

Each of you has received a consent form but I would like to remind you of a few things before we begin.

First, we will be recording the session for analysis, as soon as we finish our analysis we will delete the file, in the meantime the file will be stored in a password protected file on a password protected computer. As we write our final paper we will keep your identity anonymous.

Today we will be discussing Sexually Explicit Content, Harassment, Hate Speech, Self-harm, Illegal Activity, Misinformation, and Graphic Content. While we will be discussing sensitive content, no images will be used. The purpose of the game is to provide deeper understanding of platform governance which is inherently challenging and frustrating, and these are emotions you may feel as you play the game. If you do, we'd like you to take note of why you feel that way. With that said, if you need to take breaks or leave the game you may.

Any questions?

The Game 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. The more you are able to balance your points the more "mass appeal" your game will have which will result in more ad revenue.

For the purposes of gameplay, the game has been divided into 3 cycles, each focused on different areas of controversial content. At the start of each cycle, you will receive an investment which will allow your Trust & Safety team to make key decisions related to growth.
Throughout the game you will experience changes to both free expression and community safety points and your budget. I will track these changes in the scorecard template provided.

Round 1, Crafting your policy:
The first thing to do is decide what kind of content your company wants to allow, and what it wants to prohibit. Rather than starting from scratch, you'll select from a selection of existing policies. It is very common for new platforms to look at the incumbent social media platforms for inspiration. You'll do this by sorting the yellow policy cards into one of two categories: Allowed or Banned. As you do this, it's important to keep in mind existing legislation.

The 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 can not be held liable for content on your platform with a few exceptions (red cards, which we will discuss in a moment) and your platform can "moderate" content on the platform as long as they do not "publish" or "edit."

Now we will take a look at the "red cards" – as a small company you will do your best to avoid this content as a large legal battle may bankrupt you.

1. DMCA, Digital Millennium Copyright Act, states intermediaries, such as your platform have "indirect liability" if they do not take best faith efforts to flag copy-written material and take it down.
2. FOSTA, Fight Online Sex Trafficking Act, was passed in 2018 and outlaws content that facilitates traffickers in advertising the sale of unlawful sex acts with sex trafficking victims, if this content is your platform, you no longer have a Section 230 shield and could be sued.
3. Take down and report child porn as soon as reasonably possible after obtaining knowledge of online child sexual exploitation. This is based on federal law.

Other than this, you have the shield and the sword.

A few things to keep in mind as you take on your first role as Contentr's trust and safety team:
- You care about the company's success
- You need users to spend time on your platform in order to attract investors (you want to keep both free expression points and community safety points in a manner that keeps users returning to your site)

- You are doing your best to write clear policies that capture the wide range of content

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.

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. In this position, you'll interpret the rules on real "borderline" content, which we've drawn from real world examples. In real life, content moderators must make decisions independently and very quickly (often in as little as 3 seconds!). You won't be timed, but you should go through these as fast as you can. You may also notice that you don't have a lot of context—neither do real life content moderators, so you'll have to make the best decisions you can with the information you have. Reminder: your job is to follow the policies outlined by Contentr's policy team.

Finally, also like real life, you, as a content moderator can suggest changes to your company's trust and safety team. At the end of the cycle you will be able to add nuance to 2 existing policies. 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 how it all pans out. You started 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.

In this round you'll read the event cards, which are blue and purple. The blue cards are based on examples of pushback social media companies have faced over the years, some are based on decisions made in round 1 and round 2, some events happen regardless of the platform's efforts. The purple cards represent the results of your content moderation algorithms which were trained (rightly or wrongly) based on decisions made during round 1 and round 2.

[Bathroom break before next round]

The game is complete. We will have a short debrief. What are your initial reactions and/or takeaways in this moment?