SCENARIO EXERCISE

Content Moderation Decisions

www.aileaders-project.eu

Co-funded by
the European Union

CONTENTS

Co-funded by the European Union

Abstract

Type of OER:

SCENARIO EXERCISE

Goal/Purpose:

Puts students in the shoes of a social media company manager and shows the complexity of content-moderation decisions and how these affect the performance of the company using two videogames (Trust & Safety Tycoon and Moderator Mayhem)

Expected Learning Outcomes:

The student will be able to **analyze** the impact of content moderation and ethical decision-making on user experience and business performance.

Suggested Methodological Approach (Case-Based Learning, Problem-Based Learning...):

Problem-based learning.

Keywords:

Content moderation, stakeholder engagement, fundamental rights, ethical management, tech companies, business performance.

Introduction

This scenario exercise uses two freely accessible videogames to explore the complicated choices and trade-offs of content moderation decisions at a tech company from two points of view. The first puts the student in the shoes of a content moderator, illustrating how difficult these decisions can be and how they affect people's lives. The second puts the student in the shoes of a trust and safety manager and explores how these decisions affect the company and relationships with stakeholders.

leaders

Tools presentation

Moderator Mayhem is a videogame that will help students explore the difficult choices and trade-offs involved in content moderation activities by letting them experience the activity first-hand.

Trust and Safety Tycoon, puts the students in the shoes of a Trust and Safety Team Manager and from this point of view, explores the trade-offs and complex decisions made by a tech company that must moderate user activity.

Both games developed by Mike Masnick, Randy Lubin and Leigh Beadon at Copia and Leveraged Play.

leaders

Hands-on activities

Activity 1. Go to https://moderatormayhem.engine.is/ and just play the game! Play at least **two rounds** of **Moderator Mayhem** until you either beat the game or lose.

Activity 2. Go to https://trustandsafety.fun/ and just play the game! Play at least **one round** of **Trust and Safety Tycoon** until you either beat the game or lose.

Conclusion

By playing **Moderator Mayhem**, you now have a good idea of what content moderation is, what types of decisions need to be made and how difficult it is to make them, even if you have help from AI tools. You should also have a pretty good idea of how these decisions impact people´s lives and affect their fundamental rights.

By playing **Trusty and Safety Tycoon,** you now understand why having a defined vision for your tech company and the public you serve is very important. You also understand how variables such as user growth, ad revenue, and the content moderation activity itself have an impact on the health of your company and product. You should also have an idea of how difficult is to keep stakeholders (including your CEO) happy and how their needs and expectations do not always align and make it tough to decide what is actually the right management decision within the context of your text company.

leaders

References

Theocharis, Y., et al. (2025). Content Warning: Public Attitudes on Content Moderation and Freedom of Expression. Content Moderation Lab at TUM Think Tank. https://doi. org/10.17605/OSF.IO/F56BH

Willcox, M. (2024). Algorithmic agency and "fighting back" against discriminatory Instagram content moderation: #IWantToSeeNyome. Frontiers in Communication. , 08 January 2025 Sec. Culture and Communication, Vol. 9. https://doi.org/10.3389/fcomm.2024.1385869

Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2). https://doi.org/10.1177/2053951720943234

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951719897945

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11), 4366-4383. https://doi.org/10.1177/1461444818773059

Roberts, S. T. (2016). Commercial Content Moderation: Digital Laborers' Dirty Work. in The Intersectional Internet: Race, Sex, Class and Culture Online, edited by S. U. Noble & B. Tynes. Peter Lang. https://doi.org/10.1177/2053951720943234https://ir.lib.uwo.ca/commpub/12/?utm_source</p>

Suzor, P., et al. (2020). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation, International Journal of Communication 13. https://doi.org/10.1177/2053951720943234

Complementary material

Guidelines for instructors.

- Using this scenario exercises in class should allow you to have rich discussions about topics related to ethical management within the context of a tech company that uses human and AI-enable content moderation.

- We encourage you to direct the conversation to cover any syllabus topics you think would benefit from using these exercises. Your students will have a little bit of fun but also have opportunities to reflect on the experience.

The readings found in the **reference section** introduces several concepts relevant to the exercise. You can make them into **required readings** to prepare for the class, or they may help you prepare an introduction for students before playing the games.

leaders

Guidelines for Instructors

There are several **prompts** you can use to take the discussion in several directions. These prompts could be useful for **Moderator Mayhem.**

Have students **explain** what **content moderation** is and **reflect** upon their experience:

- Why do we think **content moderation** is important?
- How does it **affect the lives of their users**?
- What do they think of the **job itself? Share** how it felt to be in the shoes of a moderator.
- What **role** should AI play and what **potential** and **challenges** exist in deploying **automated tools** to do these tasks?
- How can we **make the right decisions?**

Guidelines for instructors

There are several **prompts** you can use to take the discussion in several directions. These prompts could be useful for **Trust and Safety Tycoon.**

Have students **share** their experience managing a **trust and safety team** and reflect upon how the **decisions** they **made** affect the health of the company:

- How did their decisions affect **user growth**, **revenues,** the **nature of the online experience** and the **public perception** of the company or **their employees**? How can they prioritize needs? Which needs **are more important?**
- How can they decide **what is the right course of action?**
- What takes priority, **shareholders** or other **stakeholders**?
- How can you balance **your bottom line** versus **societal needs?**
- What does it take to **get things right?**

.

Guidelines for instructors

Students should be able to **connect the dots between both experiences**.

You may want to prepare a **wrap up** where students **formalize their conclusions.**

For this, we suggest that you ask them to have another discussion in **smaller groups** and then **write up a list of three or four conclusions** they derived from their experiences and the class discussion **and to read them aloud to the whole class.**

This should leave them with some **food for thought**.
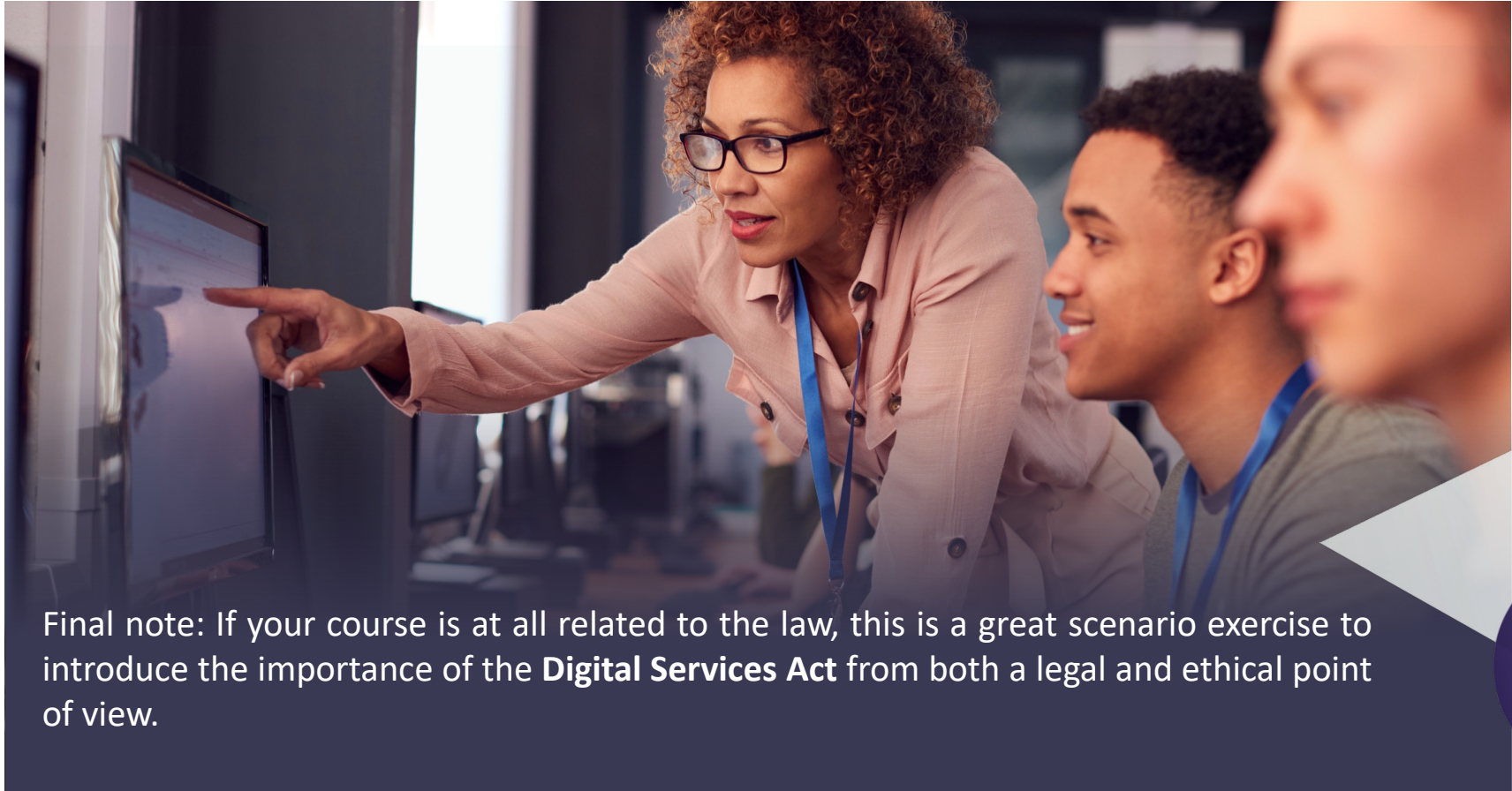
Final note: If your course is at all related to the law, this is a great scenario exercise to introduce the importance of the **Digital Services Act** from both a legal and ethical point of view.

leaders

Follow our journey

www.aileaders-project.eu

**Co-funded by
the European Union**