

Embedded Ethics
Module
Recommender System
Objectives

A hand is pointing at a screen displaying a list of GeForce GTX 750 Ti graphics cards. The list includes details such as the model name, memory size (2048 MB), and a unique identifier (e.g., 826077). The background is dark, and the text is white, creating a high-contrast, digital aesthetic.

Welcome to Embedded Ethics!

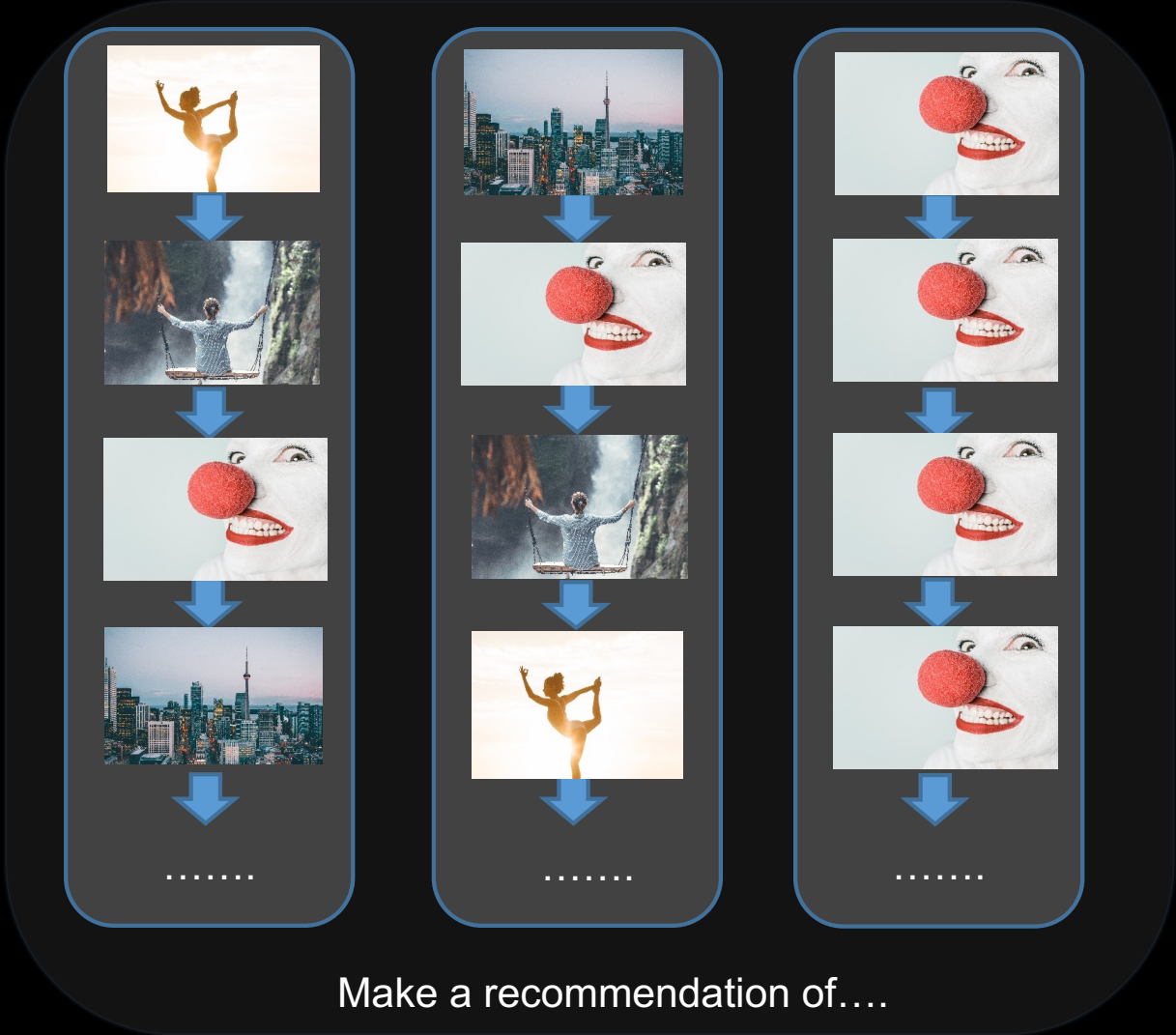
1) This is an active, participatory module – your contributions will help make it successful!

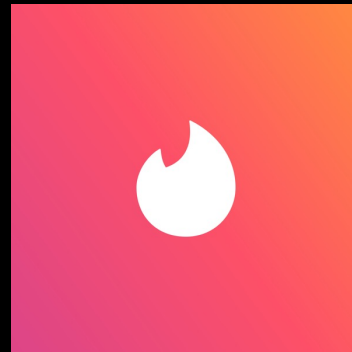
2) Our goal is not to tell you *what* to think about ethical problems, but give you some tools for *how* to think about them.

Objective function of a recommender system

To maximize { Clicks
Viewing time
Engagement
Logins
Etc }

Make a recommendation of....





Recommender systems are diverse!

Warmup (Group) Exercise – Pen and Paper (6-8 minutes)

- What objective is each intended to maximize?
 1. Amazon
 2. Netflix
 3. Duolingo (language learning app)
 4. Tinder
 5. Facebook

Sometimes recommender systems can narrow down content for users.

When there are a lot of choices, what are the alternatives to recommender systems?



Going through large amounts of content by yourself



Relying on expertise of others



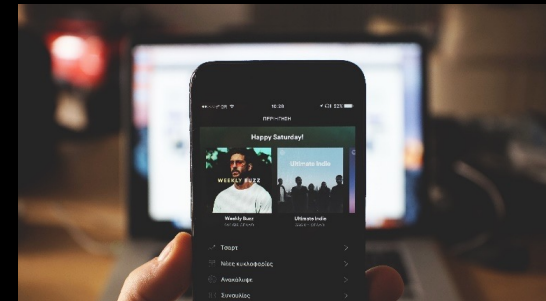
Random chance

Discussion Question

How could you improve collaborative filtering to decrease social convergence, echo chambers, etc?



Sometimes recommender systems can also help users find content that is appropriate for them (which they couldn't find easily just by previewing it):



Positive Values that can be Promoted by Recommender Systems



Happiness/well-being

(Both directly and by making lives more
efficient)



Autonomy

(Increasing our control over our
own lives)

Two Kinds of Negative Moral Impacts



Negative consequences

Loss of happiness, causing pain, etc

Violations of rights

If someone has a right that you shouldn't do X to them, you shouldn't do X, even if it produces the best consequences)

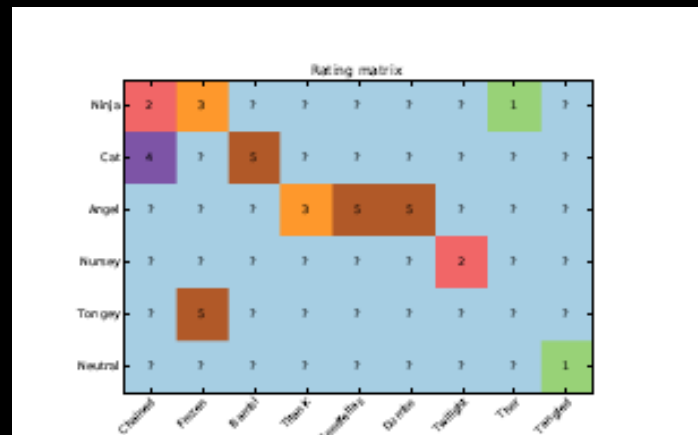
Not just legal, but also moral!

Some rights can be waived (by voluntary agreement, you make it OK to do X).



Part 1:
Collaborative Filtering
and Social Convergence

Collaborative filtering “uses the known preferences of a group of users to make recommendations or predictions of the unknown preferences for other users.” (Su and Khoshgoftaar, 2009)



Social Convergence: Many recommender systems choose for you based on what other people who make similar choices have chosen.

Poll Questions

- In your opinion, to what extent is social convergence a problem in each of the following apps?
 1. Duolingo Language Practice Sets
 2. Tinder
 3. Netflix
 4. Facebook

Echo chamber: an environment where a person encounters only information or views that reflect and reinforce their own information or views.

They “may limit the exposure to diverse perspectives and favor the formation of groups of like-minded users framing and reinforcing a shared narrative.”
(Cinelli et al 2021)



Group Exercise 2

- Suppose that you are an intern at Reddit in charge of the site's recommender system.
- This is a vast oversimplification, but imagine that the algorithm works this way: users subscribe to subreddits, and see the posts that are most upvoted by users of those subreddits. Advertisements are sprinkled occasionally into the posts.
- Now imagine that Reddit has just been acquired by a billionaire who has fired half of the employees and demanded that as part of "Reddit 2.0", the algorithm must be improved to create as much engagement and profit as possible.
- List at least three changes you would make. We'll discuss them in 6-8 minutes.

Discussion Question

Most problems with social similarity, echo chambers, etc, are about **negative consequences**.

Can you think of any ways that recommender systems might **violate peoples rights**?



Part 2: Manipulation

One way that a recommender systems might violate someone's rights is by **manipulating** them.

To get a better grasp of manipulation, let's see some examples of it.



Conditioning is an attempt to get someone to adopt a pattern of behaviour by rewarding or punishing their actions.





A **guilt trip** is using an inappropriate amount of guilt to influence someone to do something.

Discussion Question

A **definition** of manipulation would explain what all of these cases have in common with each other.

What do the previous three examples have in common with each other that make them count as 'manipulation'?

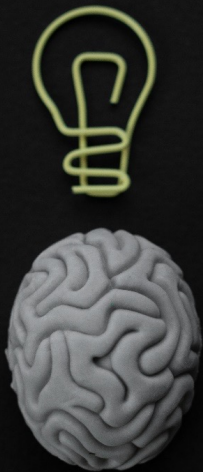


What do these actions have in common with each other that make them count as 'manipulation'?

One theory: *“manipulative action is the intentional attempt to get someone's **beliefs, desires, or emotions** to violate their norms or ideals, from the perspective of the manipulator.”*

(Robert Noggle, “Manipulative Actions: A Conceptual and Moral Analysis”)





A standard for beliefs:

"Believe only the truth."

Deception is a kind of manipulation.



A standard for desires:

“Desire only what you judge that you have reason to desire.”

Creating an **addiction** in someone is a form of manipulation



Standards for emotions:

“Base your emotions on true beliefs.”

“Ensure that emotions highlight only things that are genuinely relevant to your deliberations.”

The Facebook Papers (2021)



In October 2021, a number of internal Facebook documents were made public by a whistleblower named Frances Haugen.

Many thought that these documents showed that Facebook was aware of many of the ethically dubious consequences of their social media platforms.

from the files

Weight Decision 12/15/2017

Component	Final Weight for 2018Q1
Like	1
Reaction, Reshare without Text	5
Non-sig Comment, Non-sig Reshare Non-sig Message, Rsvp	15
Significant Comment, Significant Reshare, Significant Message	30
Groups Multiplier (Non-friends)	0.5
Strangers Multiplier (non-friend-of-friend, small pages)	0.3

Wall Street Journal, "Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead"

- In deciding which posts to present to users, Facebook has an explicit formula describing the relative weights of certain factors.
- Facebook introduced this formula in order to drive more meaningful interactions.
- "The goal of the algorithm change was to reverse the decline in comments, and other forms of engagement, and to encourage more original posting. It would reward posts that garnered more comments and emotion emojis, which were viewed as more meaningful than likes, the documents show."



“While the FB platform offers people the opportunity to connect, share and engage, an unfortunate side effect is that harmful and misinformative content can go viral, often before we can catch it and mitigate its effects,” he wrote. “Political operatives and publishers tell us that they rely more on negativity and sensationalism for distribution due to recent algorithmic changes that favor reshares.” (Internal Facebook Memo, quoted by the *Wall Street Journal*)

Poll Question

Which of the following best describes your reaction to the amplification of angry content in the Facebook Papers?

1. It is ethically permissible
2. It is unethical, primarily because of its bad consequences (polarization, violence, etc)
3. It is unethical, primarily because it is manipulative.
4. It is unethical, primarily because of some other reason.

Part 3: Trust

Wordcloud Exercise

- Think of a recommender system that you trust. Which features make you trust it?



Giving users control over what they see in their feeds may also lead to more (justified) trust. (Stray, “Beyond Engagement”)

E.g. ‘see less often’ or ‘hide post’ functions in feeds

Discussion Question

What sort of personal controls would you want to have over your feeds in the social media platforms you use?





Some software designers have even proposed changing the **objective function** of many recommender systems: instead of maximizing engagement, they should maximize well-being. (Stray, “Beyond Engagement”)

Conferences > 2020 IEEE International Confe... ?

IEEE 7010: A New Standard for Assessing the Well-being Implications of Artificial Intelligence

Publisher: **IEEE**

[Cite This](#)

[PDF](#)

Daniel Schiff ; Aladdin Ayesh ; Laura Musikanski ; John C. Havens [All Authors](#)

Summary

- Recommender systems are powerful and valuable to individuals and society.
- It is important to be conscious of the harms and violations of rights they might cause.
- One example of a possible harm is social convergence or echo chambers.
- One example of a violation of rights is manipulation.
- If you have questions or thoughts, I'm happy to chat more offline – steven.coyne@mail.utoronto.ca



Acknowledgements

This module was created as part of an Embedded Ethics Education Initiative (E3I), a joint project between the Department of Computer Science¹ and the Schwartz Reisman Institute for Technology and Society², University of Toronto.

Instructional Team:

Roger Grosse, Steven Coyne, Emma McClure, Rahul Krishnan

Faculty Advisors:

Diane Horton¹, David Liu¹, and Sheila McIlraith^{1,2}

Department of Computer Science
Schwartz Reisman Institute for Technology and Society
University of Toronto

References

- Noggle, Robert. “Manipulative Actions: A Conceptual and Moral Analysis”, *American Philosophical Quarterly* 33(1), p.43-55
- Russell, Stuart. *Human Compatible*. New York: Viking Press, 2019
- Stray, Jonathan. “Beyond Engagement” Accessed online: <https://partnershiponai.org/beyond-engagement-aligning-algorithmic-recommendations-with-prosocial-goals/>
- Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." *Advances in artificial intelligence* 2009 (2009).