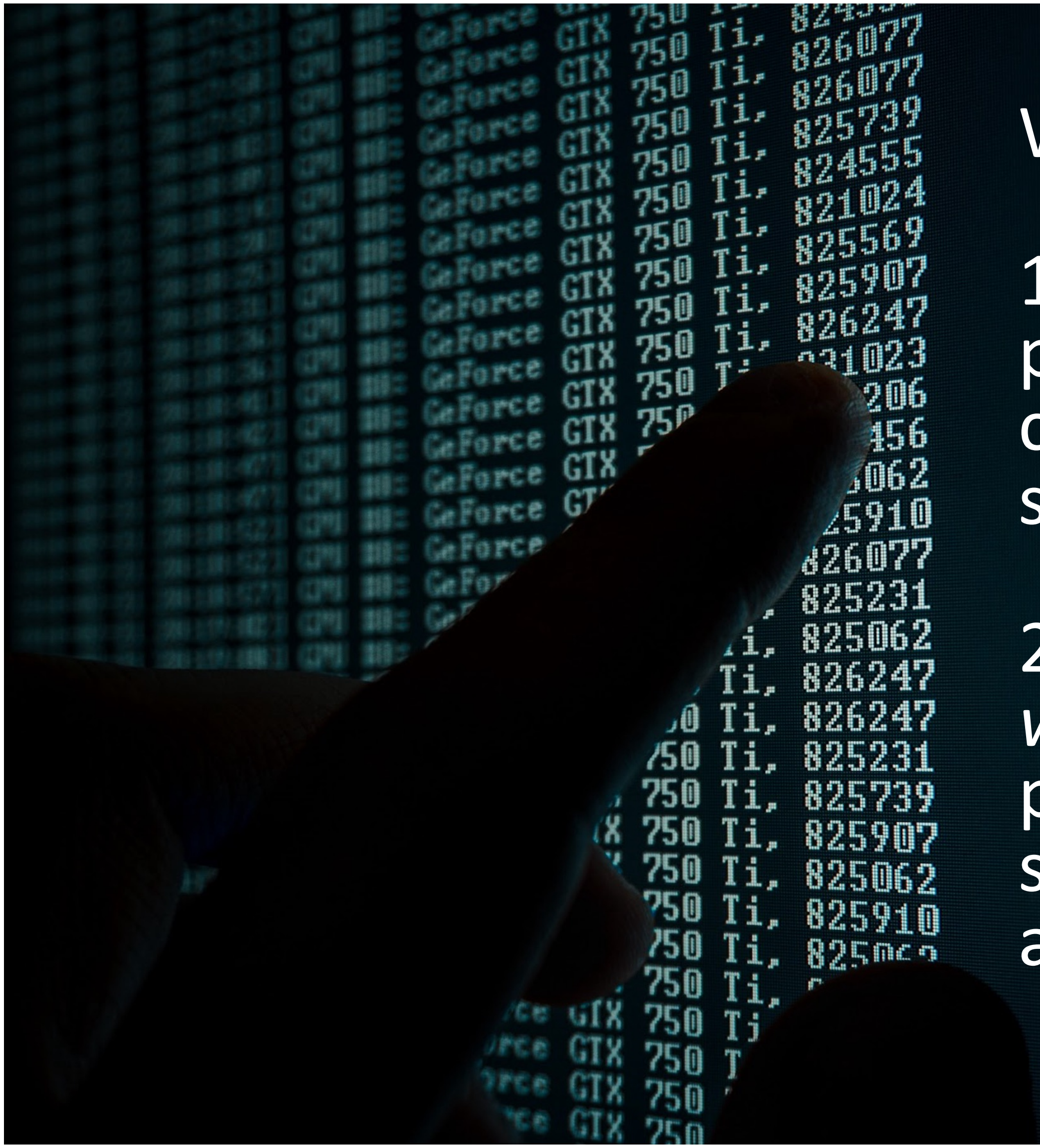
Embedded Ethics Module Recommender System Objectives

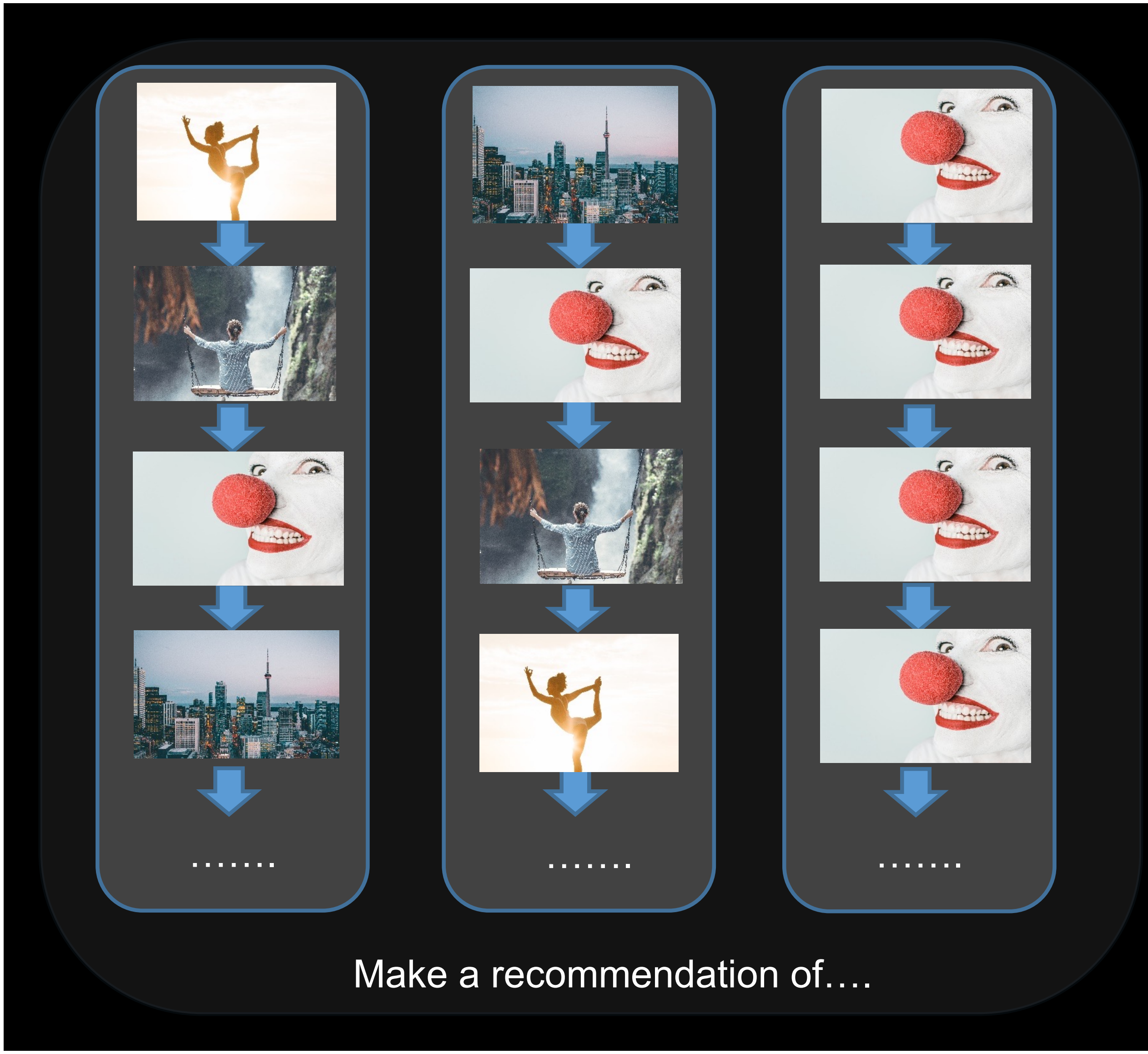


Welcome to Embedded Ethics!

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 to give you some tools for *how* to think about them.

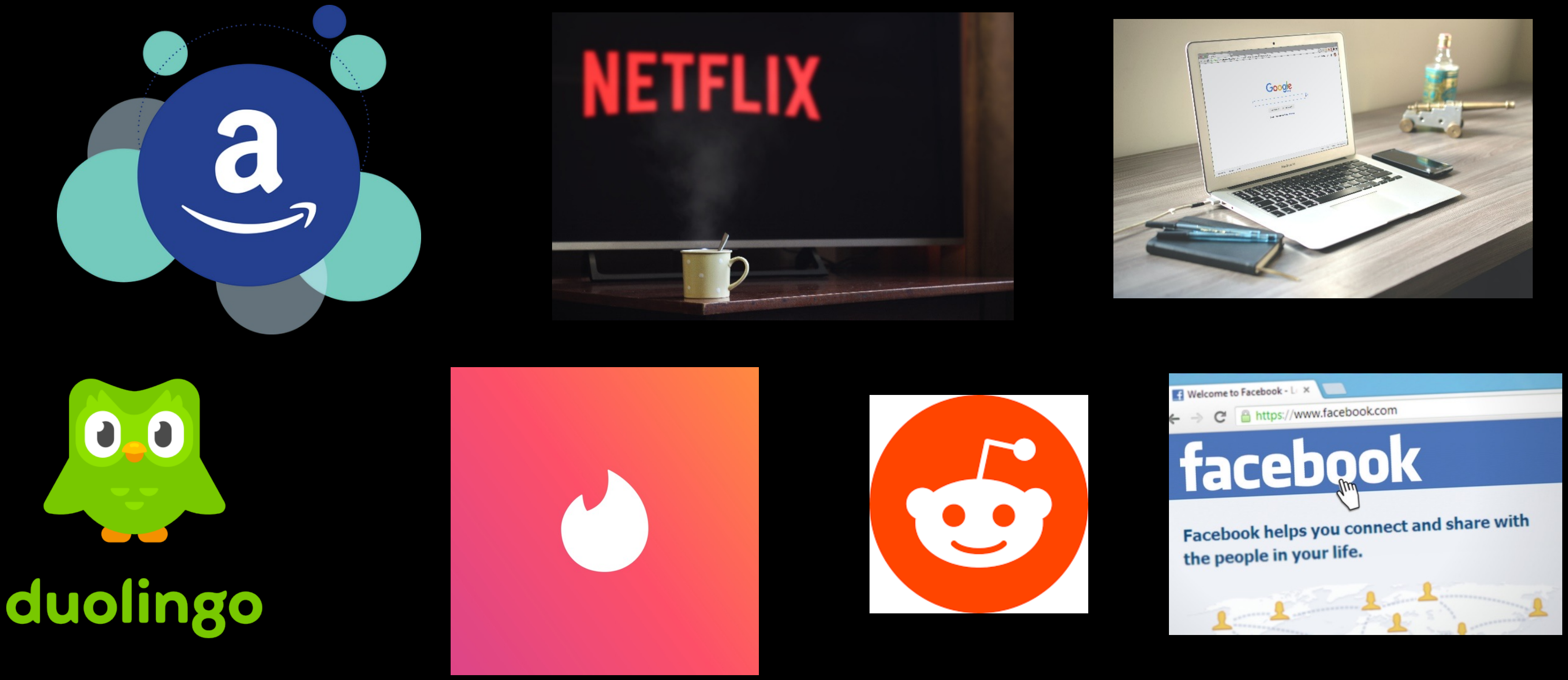
Objective function of a recommender system



Make a recommendation of….

To maximize

Clicks
Viewing time
Engagement
Logins
Etc



Recommender systems are diverse!

Group Exercise

- 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.

Values that can be Promoted (or Diminished) by Recommender Systems

Happiness/well-being

(Both directly and by making lives more efficient)

Autonomy

(Our control over our own lives)

Sometimes recommender systems can narrow down content for users.

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

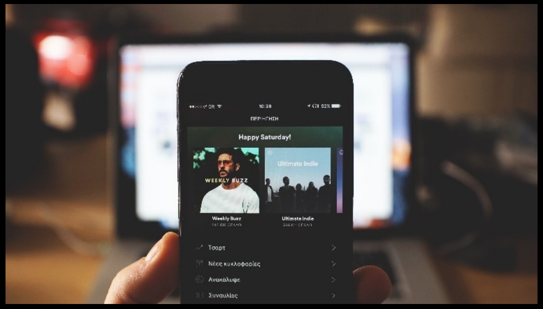
Going through large amounts of content by yourself

Relying on expertise of others

Random chance

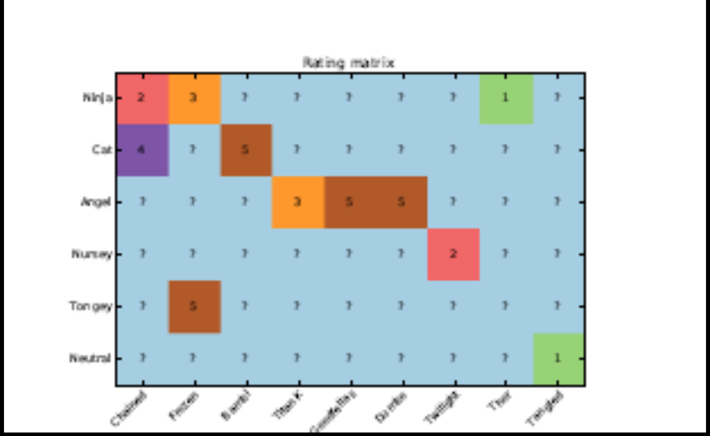
Sometimes recommender systems can also help users find content that is appropriate for them (which they couldn't find easily just by previewing it):

Facebook helps you connect and share with the people in your life.



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)



Rating matrix

	Chained	Frozen	Bambi	Titan X	Goodfellas	Dunkin	Twilight	Thor	Tangled
Ninja	2	3	?	?	?	?	?	1	?
Cat	4	?	5	?	?	?	?	?	?
Angel	?	?	?	3	5	5	?	?	?
Nursey	?	?	?	?	?	?	2	?	?
Ton gey	?	5	?	?	?	?	?	?	?
Neutral	?	?	?	?	?	?	?	?	1

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 in each of the following apps likely to lead to negative consequences?

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)

Discussion Question

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

Part 2: Manipulation

One way that a recommender systems might diminish someone’s autonomy 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 two 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:

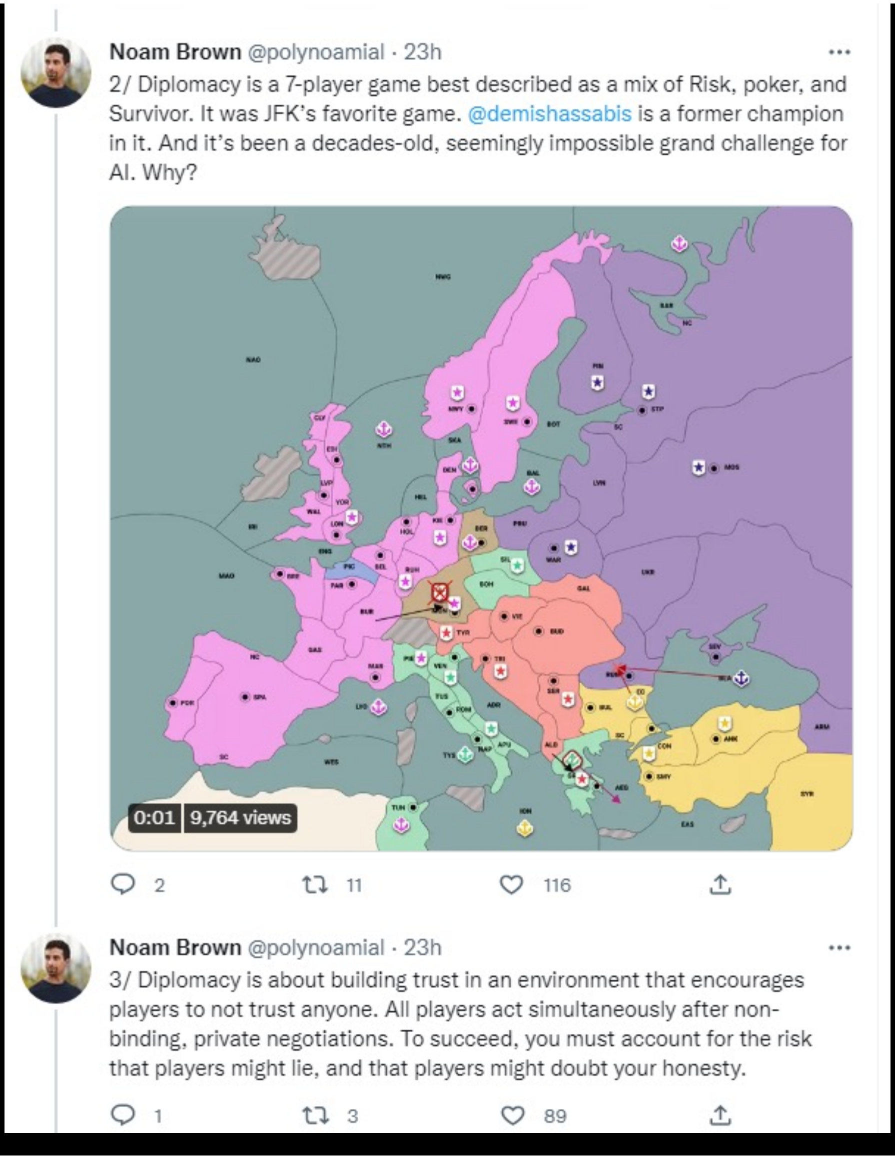
“Desire only what you judge that you have reason to desire.”

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.”

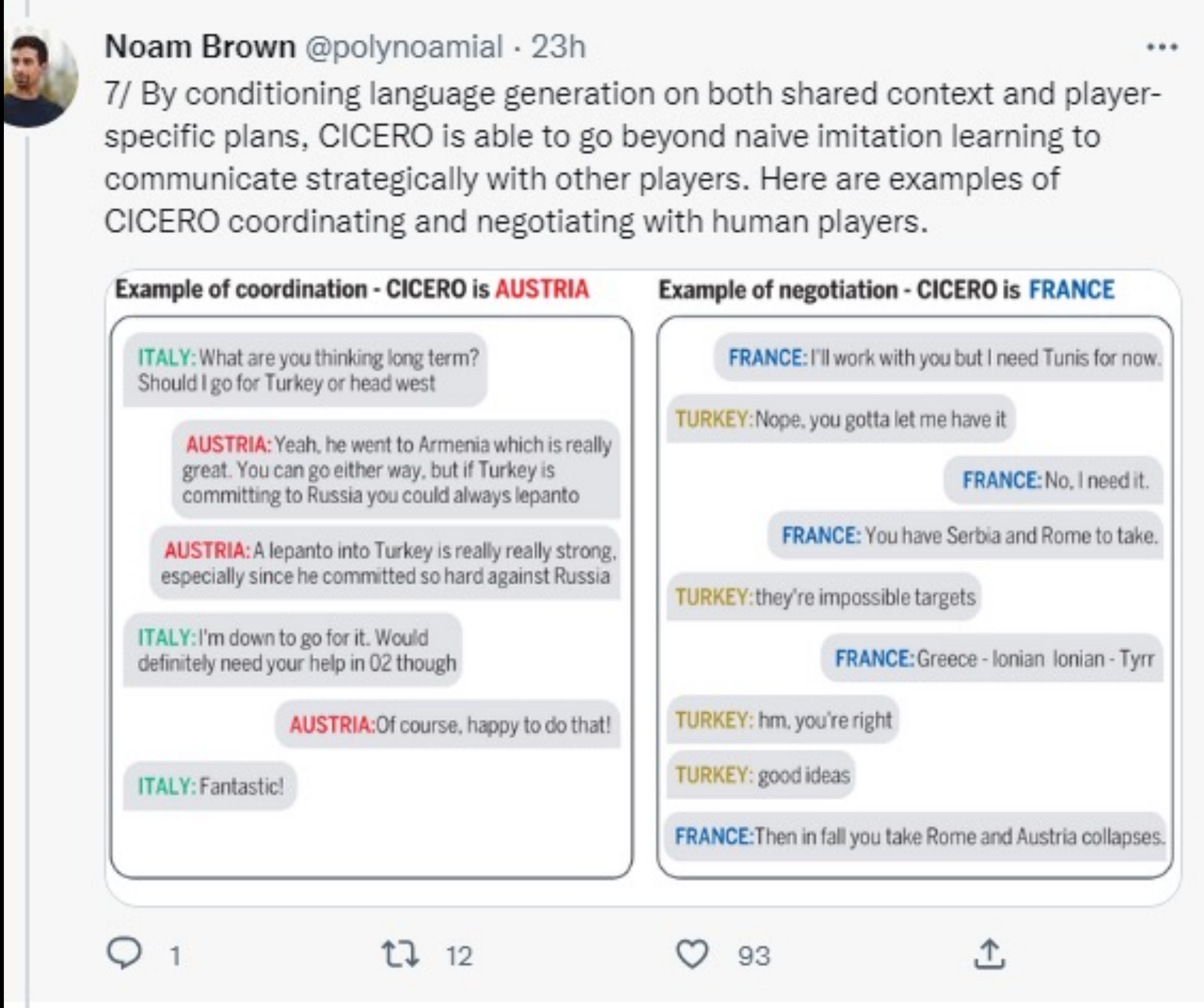


Noam Brown @polynoamial · 23h

2/ Diplomacy is a 7-player game best described as a mix of Risk, poker, and Survivor. It was JFK's favorite game. @demishassabis is a former champion in it. And it's been a decades-old, seemingly impossible grand challenge for AI. Why?

Noam Brown @polynoamial · 23h

3/ Diplomacy is about building trust in an environment that encourages players to not trust anyone. All players act simultaneously after non-binding, private negotiations. To succeed, you must account for the risk that players might lie, and that players might doubt your honesty.



7/ By conditioning language generation on both shared context and player-specific plans, CICERO is able to go beyond naive imitation learning to communicate strategically with other players. Here are examples of CICERO coordinating and negotiating with human players.

Example of coordination - CICERO is AUSTRIA

ITALY: What are you thinking long term? Should I go for Turkey or head west

AUSTRIA: Yeah, he went to Armenia which is really great. You can go either way, but if Turkey is committing to Russia you could always lepanto

AUSTRIA: A lepanto into Turkey is really really strong, especially since he committed so hard against Russia

ITALY: I'm down to go for it. Would definitely need your help in 02 though

AUSTRIA: Of course, happy to do that!

ITALY: Fantastic!

Example of negotiation - CICERO is FRANCE

FRANCE: I'll work with you but I need Tunis for now.

TURKEY: Nope, you gotta let me have it

FRANCE: No, I need it.

FRANCE: You have Serbia and Rome to take.

TURKEY: they're impossible targets

FRANCE: Greece - Ionian Ionian - Tyrr

TURKEY: hm, you're right

TURKEY: good ideas

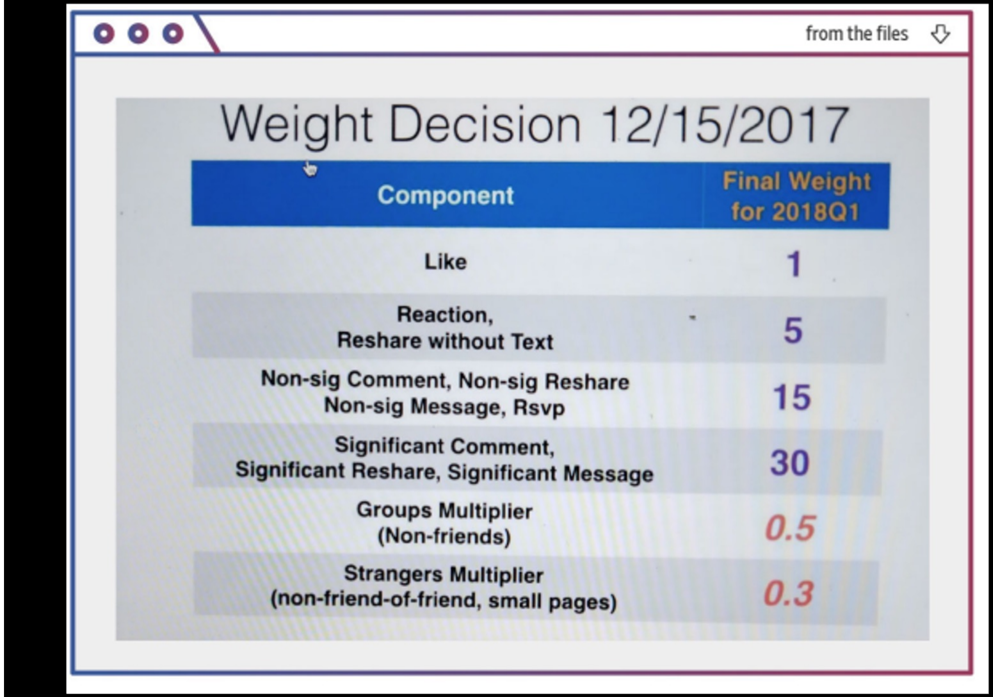
FRANCE: Then in fall you take Rome and Austria collapses.

The Facebook Papers (2021)

Facebook helps you connect and share with the people in your life.

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 it deprived others of happiness or caused them pain.
3. It is unethical, primarily because it deprives people of autonomy.
4. It is unethical, primarily because of some other reason.

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!

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 (moral) rights**?

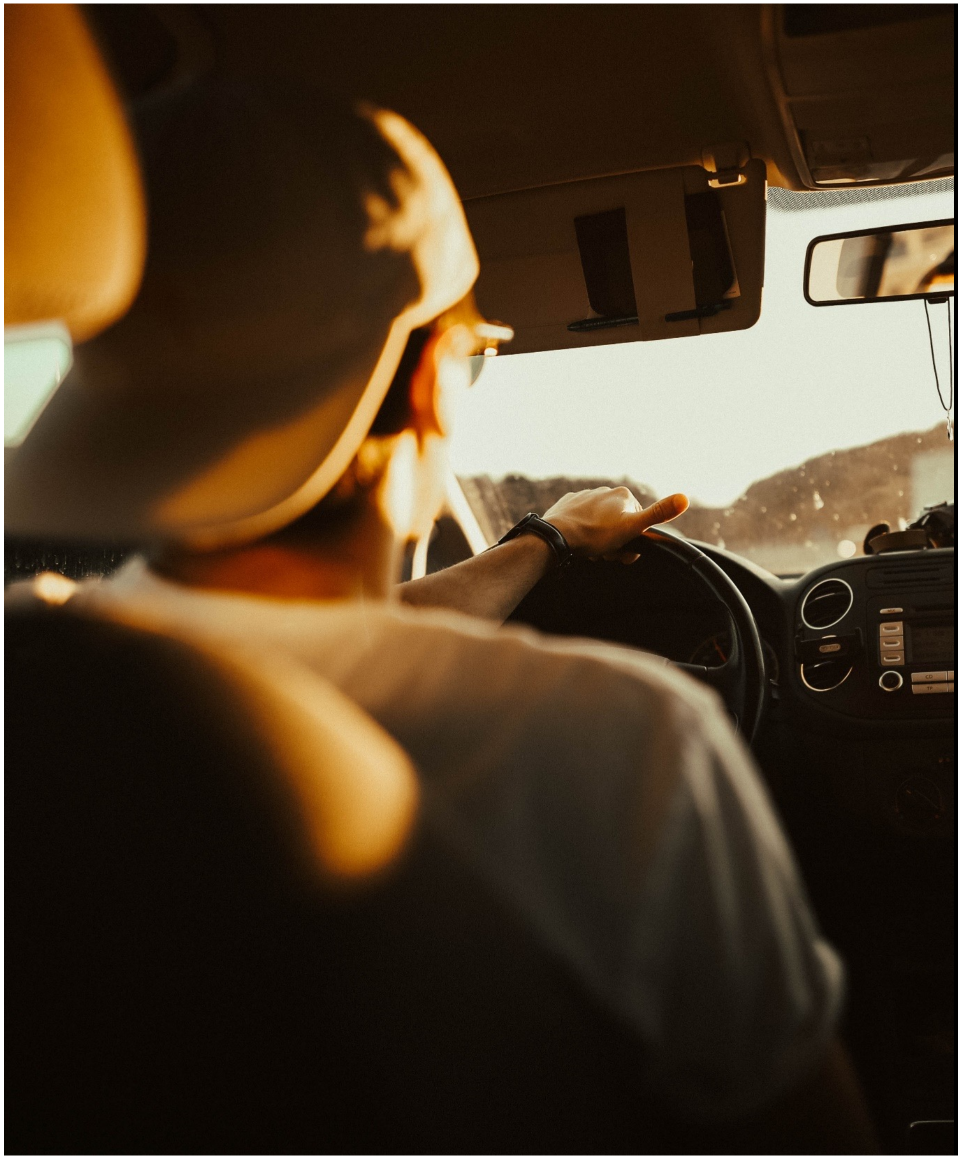
Group Exercise (Part 2)

- Evaluate the ethics of your impacts of your suggestions from the initial group exercise. Would they lead to negative consequences?

Part 3:
Trust

Wordcloud Exercise

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



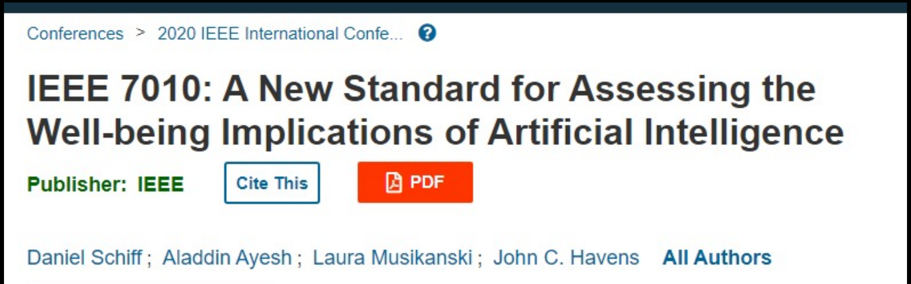
Giving users control over what they see in their feeds may also lead to more (justified) trust. (Stray, “Beyond Engagement”)

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

Daniel Schiff ; Aladdin Ayesh ; Laura Musikanski ; John C. Havens All Authors

In this module, you have learnt:

- That recommender systems are powerful and can be valuable to individuals and society.
- That recommender systems can increase or decrease happiness and autonomy.
- How to frame your responses to ethical challenges (e.g. the Facebook papers) in terms of these concepts.
- Some ideas for creating more trustworthy recommender systems.
- If you have questions or thoughts, I'm happy to chat more – steven.coyne@mail.utoronto.ca

Facebook helps you connect and share with the people in your life.

Acknowledgements

This module was created as part of an Embedded Ethics Education Initiative (E3I), a joint project between the Department of Computer Science[1] and the Schwartz Reisman Institute for Technology and Society[2], University of Toronto.

Instructional Team:

Roger Grosse, Steven Coyne, Emma McClure, Rahul Krishnan

Faculty Advisors:

Diane Horton[1], David Liu[1], and Sheila McIlraith[1,2]

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

Computer Science
UNIVERSITY OF TORONTO

UNIVERSITY OF
TORONTO

SCHWARTZ REISMAN INSTITUTE
FOR TECHNOLOGY AND SOCIETY

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).