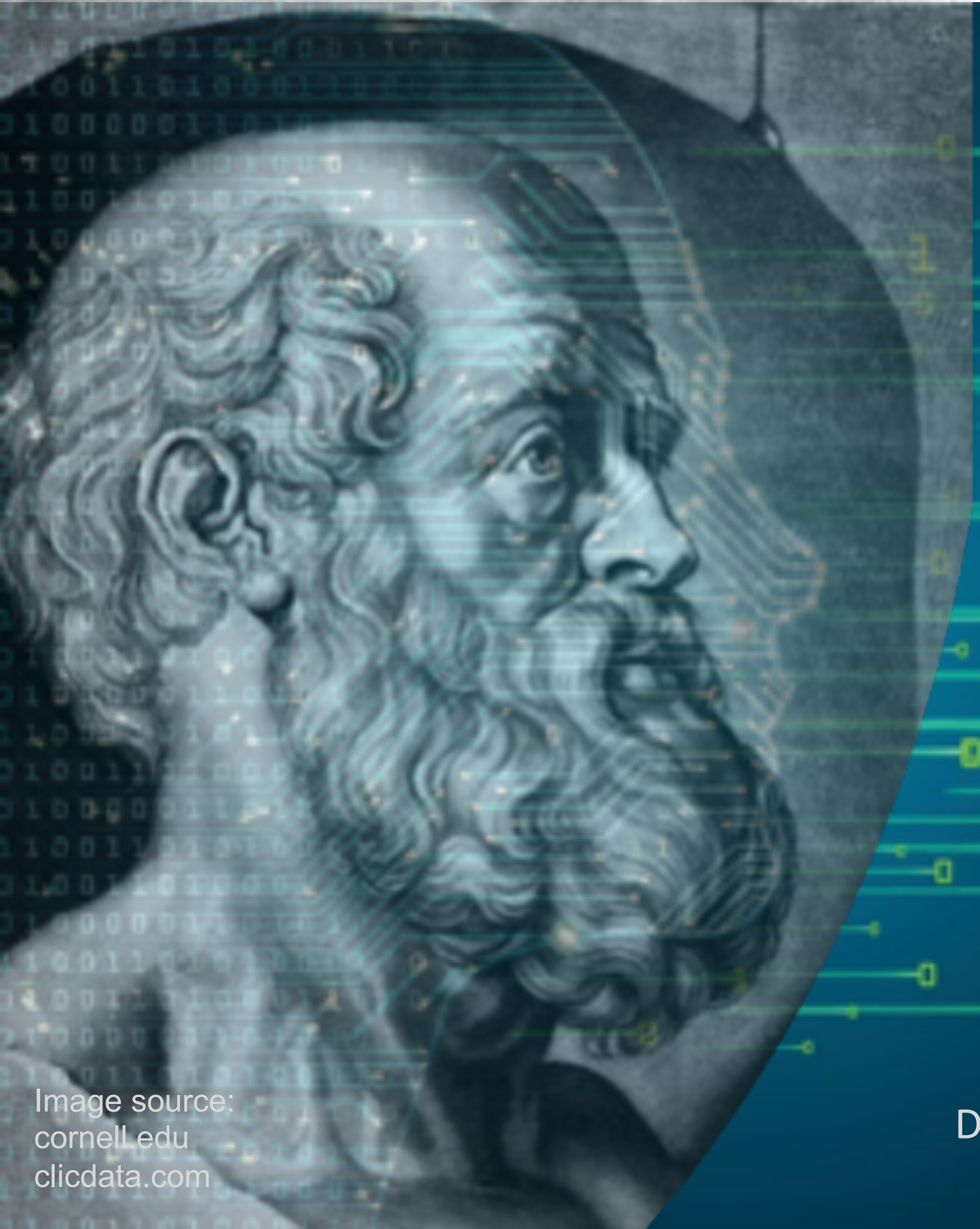


Image source:
cornell.edu
clicdata.com

CSC311 Embedded Ethics Module

Fall 2021

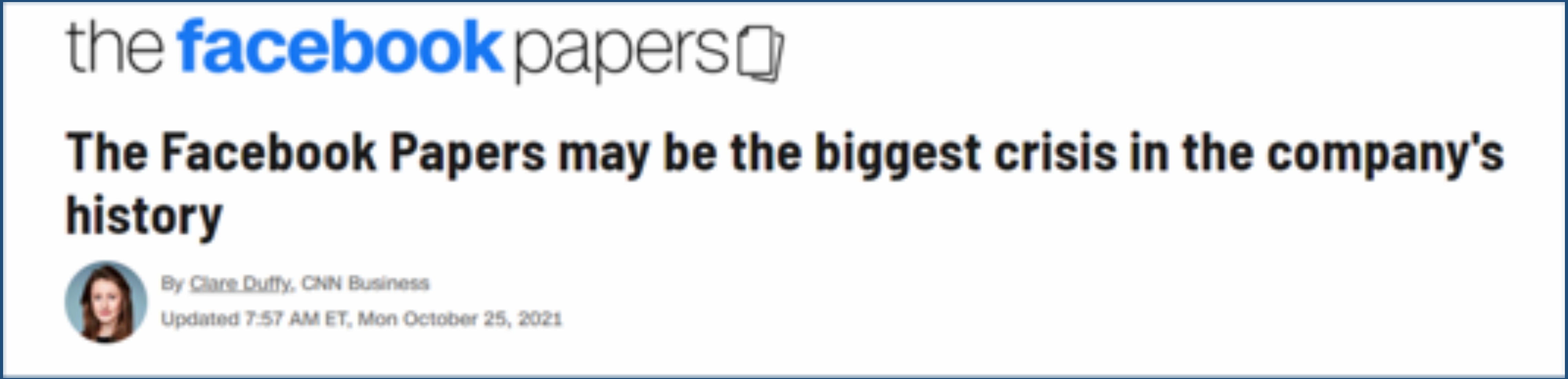
University of Toronto

Department of Computer Science and Schwartz Reisman Institute

Introduction to Embedded Ethics (Online)

- The goal of this module is not to tell you what to think about ethical issues, but to make you more comfortable in identifying and discussing them.
- Feel free to use the chat! We want to hear your thoughts and reactions, and will do our best to respond to a few of them in real time.
- Be respectful, but don't hesitate to disagree with one another. In the chat and breakout groups, address your comments to the person's views or arguments, rather than the person themselves.

The Facebook Papers



the facebook papers

The Facebook Papers may be the biggest crisis in the company's history

By Clare Duffy, CNN Business

Updated 7:57 AM ET, Mon October 25, 2021

Source: https://edition.cnn.com/2021/10/25/tech/facebook-papers/index.html

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

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

Source: bloomberg.com

The Ethics of Recommender Systems

- Our goal will be to introduce you to some ethical concepts that might be relevant to thinking about what lessons can be drawn from the Facebook Papers – and about the ethics of recommender systems in general.
- Note that this is different from the **legality** of these issues!

Issue 1: Content Moderation

“Facebook researchers documented how its platform has contributed to divisive, inter-religious conflict in India, according to internal records.

Employees flagged that human traffickers in the Middle East used the site to lure women into abusive employment situations. They warned that armed groups in Ethiopia used the site to incite violence against ethnic minorities. They sent alerts to their bosses about organ selling, pornography and government action against political dissent, according to the documents. They also show the company’s response, which in many instances is inadequate or nothing at all.” (Source: Wall Street Journal, https://www.wsj.com/articles/the-facebook-files-11631713039)

Issue 1: Content Moderation

Choose which of these
to present to the user

To maximize

Clicks
Viewing time
Engagement
Logins
Etc.

Issue 1: Content Moderation

Choose which of these to present to the user, when to present them, in what order….

To maximize

Clicks
Viewing time
Engagement
Logins
Etc.

- The problem of **content moderation:** are there any posts that should not be presented under any circumstances?

Discussion Question

Instructions for building an untraceable assault rifle

False information claiming that COVID vaccines cause autism

A call to “take back” Parliament through force

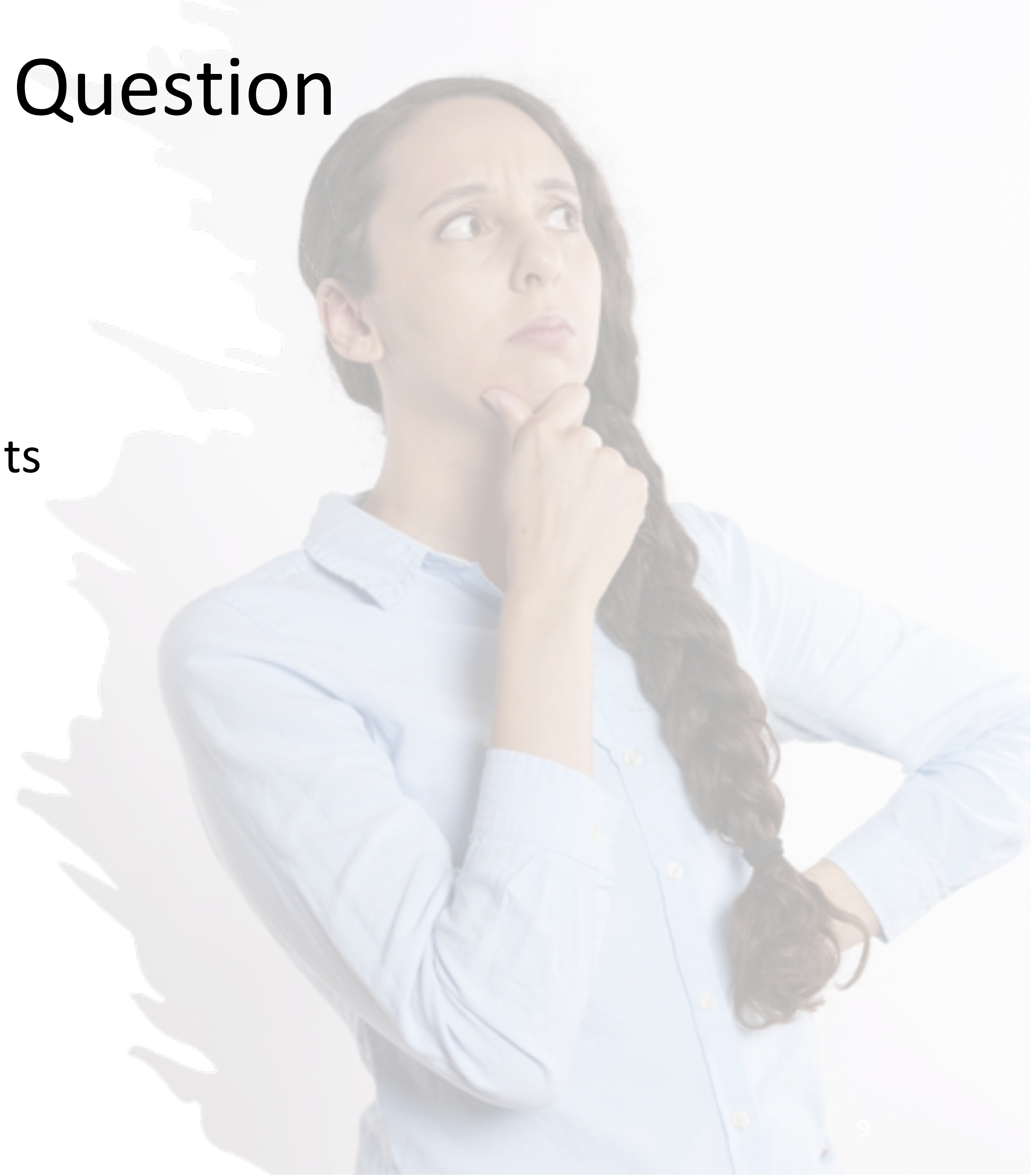
Which of these posts, if any, should be prohibited on a social media platform?

Mentimeter poll

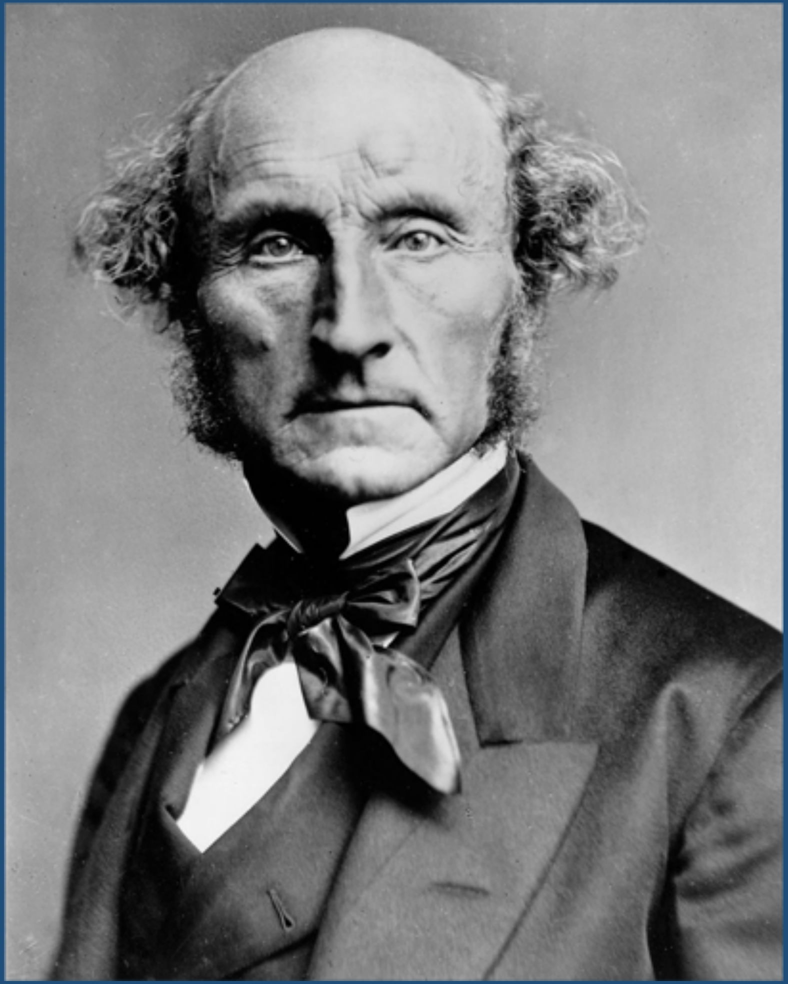
Discussion Question

In the chat (or raise your hand to speak):

What **principle** (if any) guided your judgments about which content to prohibit?



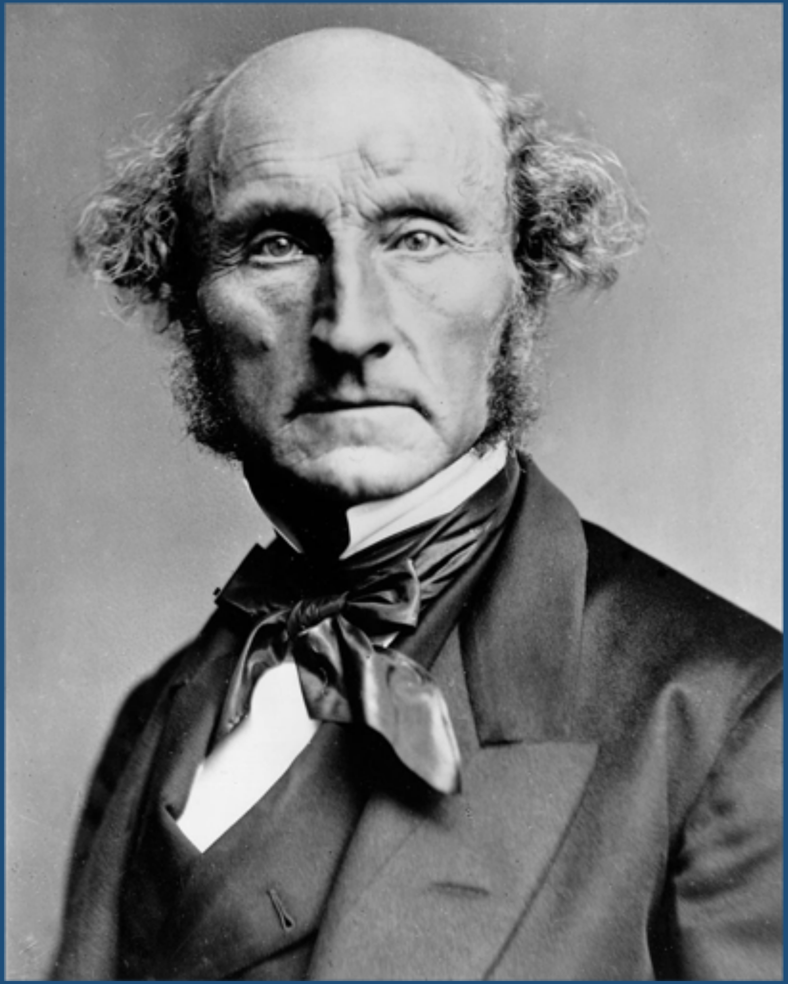
Issue 1: Content Moderation



Mill’s Harm Principle

“The object of this Essay is to assert one very simple principle, as entitled to govern absolutely the dealings of society with the individual in the way of compulsion and control, whether the means used be physical force in the form of legal penalties, or the moral coercion of public opinion. That principle is, that the sole end for which mankind are warranted, individually or collectively, in interfering with the liberty of action of any of their number, is self-protection. That the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others.” (Mill, *On Liberty*)

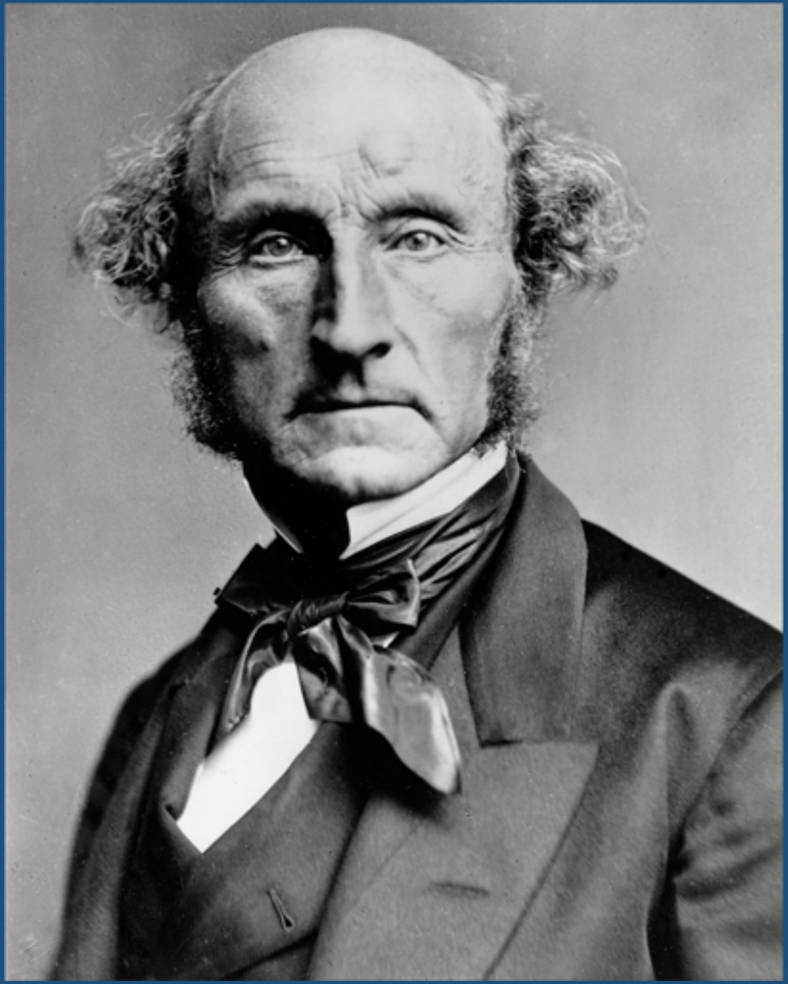
Issue 1: Content Moderation



Mill thought that the Harm Principle implied very strong protections for speech:

> "If all mankind minus one were of one opinion, and only one person were of the contrary opinion, mankind would be no more justified in silencing that one person, than he, if he had the power, would be justified in silencing mankind." (Mill, 18)

Issue 1: Content Moderation

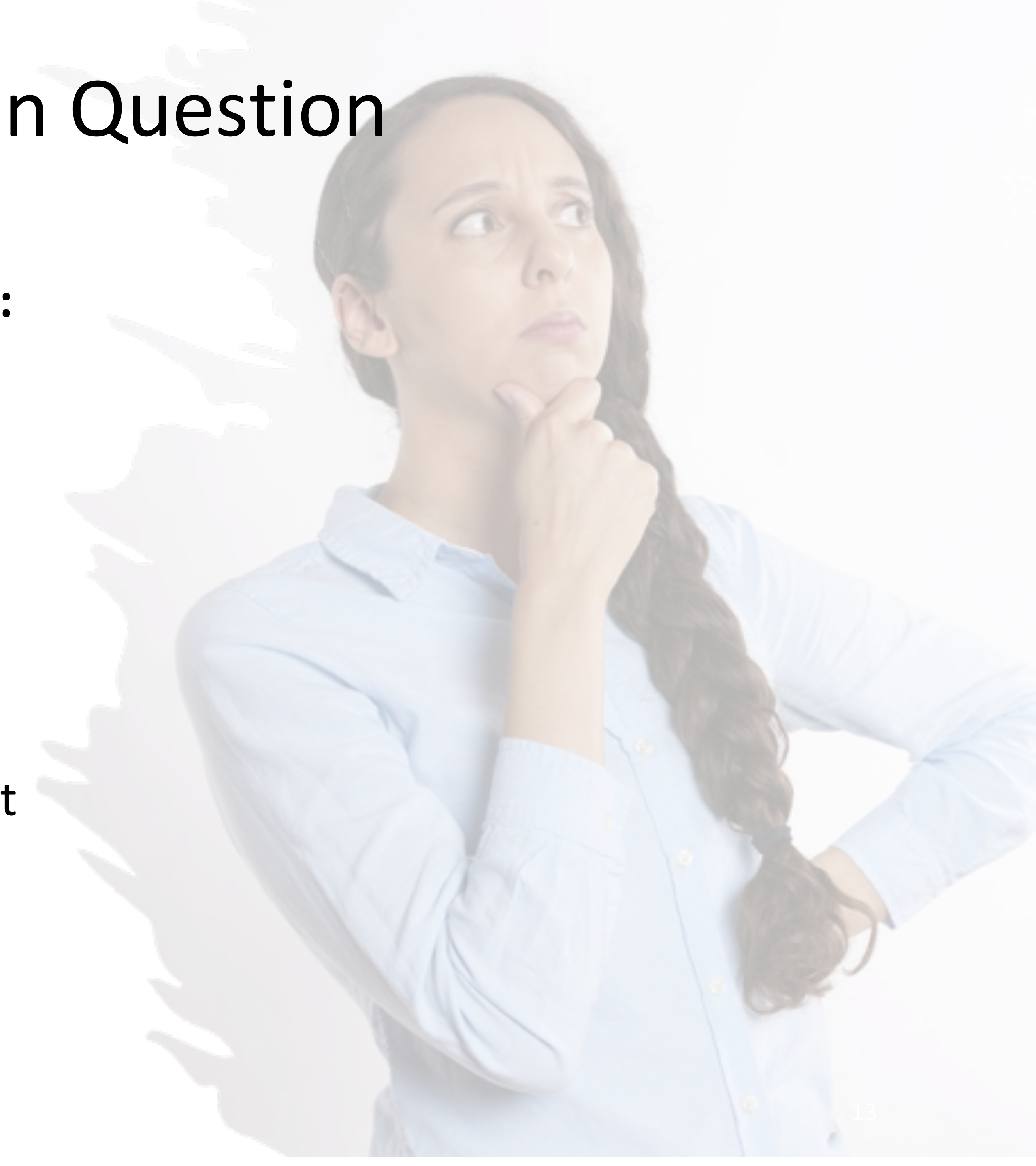


Mill's argument for this conclusion rests on the importance of free speech and debate in coming to know the truth:

> "If the opinion is right, they are deprived of the opportunity of exchanging error for truth: if wrong, they lose, what is almost as great a benefit, the clearer perception and livelier impression of truth, produced by its collision with error." (Mill, 18)

He thinks that speech should only be prohibited in cases where it results in a direct harm to a specific person (e.g. inciting a crowd to harm them, or lying about them in a way that harms their reputation)

Discussion Question



In the chat (or raise your hand to speak):

Going back to our three examples:

1. Gun manufacturing instructions
2. False COVID information
3. A call to “take back” Parliament

How compelling do you find Mill’s argument for each of these three examples?

Discussion Question

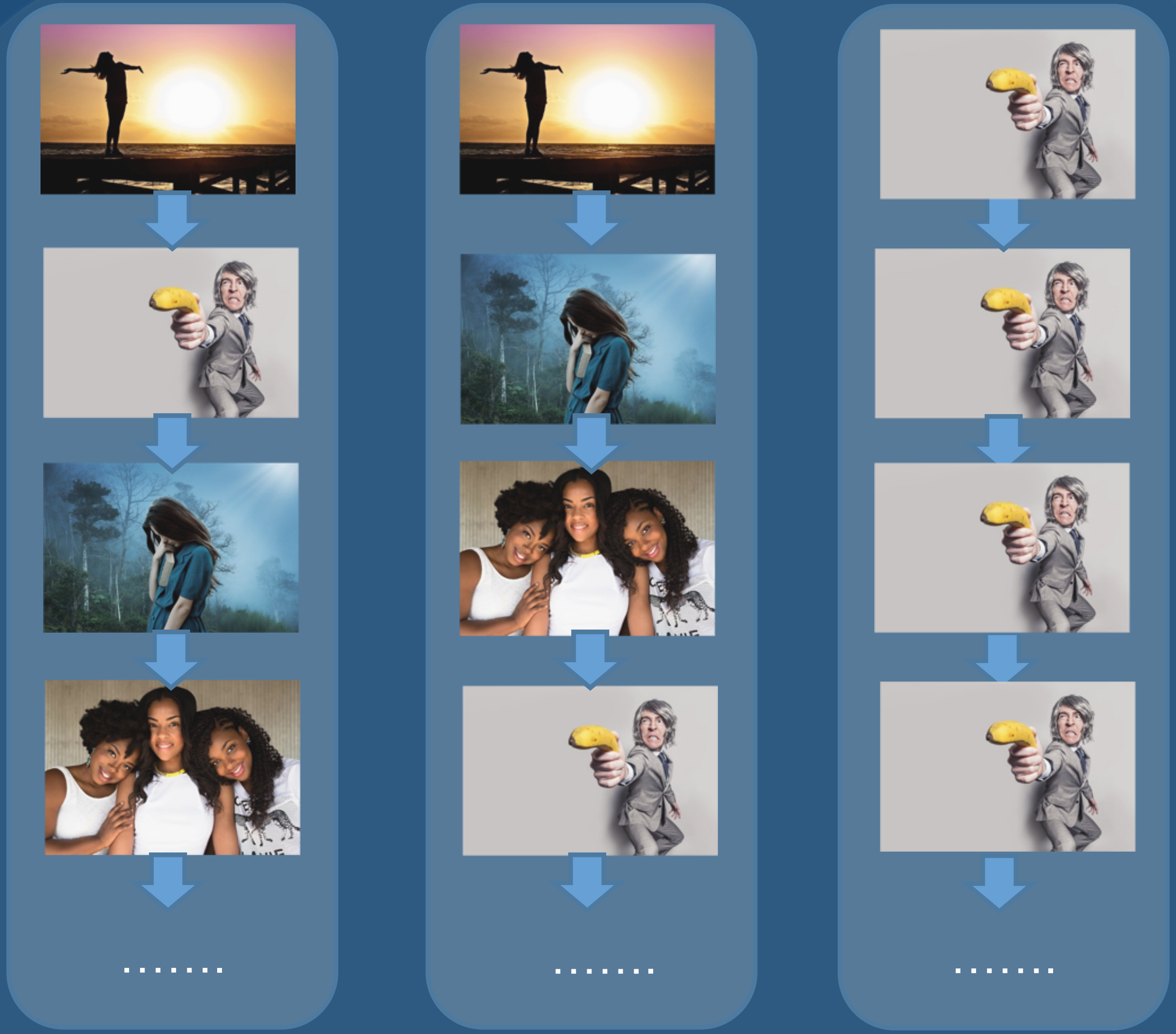
In the chat (or raise your hand to speak):

Let’s now assume that there are no failures of content moderation.

What other ethical issues might arise with recommender systems?

Issue 2: The Feed

- Recommender systems don’t just give users the opportunity to look up certain content they are interested in.
- They also choose which content to **amplify** (by posting it more prominently or more often).

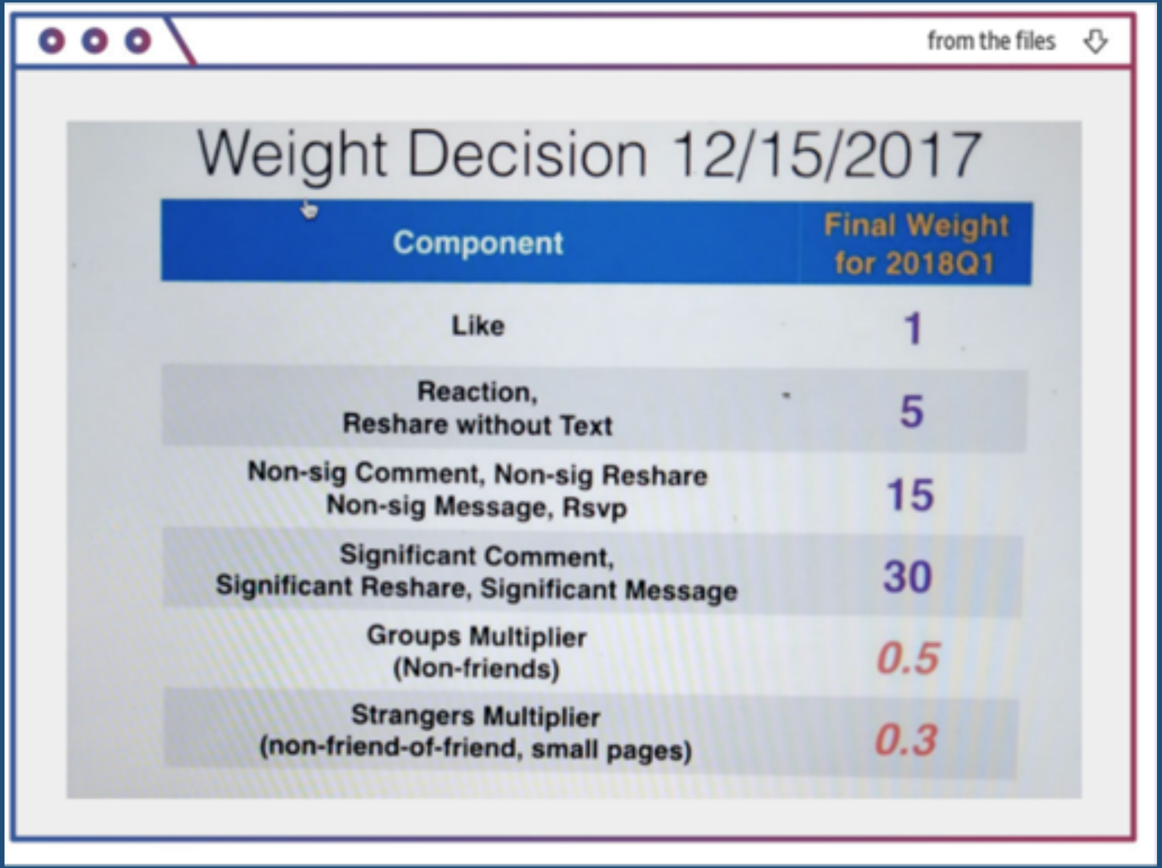


Choose between….

To maximize

Clicks
Viewing time
Engagement
Logins
Etc

Issue 2: The Feed



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.”

Issue 2: The Feed

"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*)

Breakout Group Exercise

We will look at four cases showing different ways that social media platforms might influence their users.

For each case presented:

1) Discuss each case as a group. Rank them from most ethical to least ethical. Submit one answer per group via google form, attempting to capture the consensus in your group. [10 minutes]

   Google LINK:

1) We will then take a vote on how unethical you think the influence is, on a scale from 1 (raises no ethical problems) to 5 (very unethical, to the point that it should be legally prohibited).

Mentimeter

Breakout Group Exercise

Examples:

1. A recommender algorithm that presents users with advertisements similar to the posts they have clicked on in the past.
2. A recommender system that presents advertisements for expensive luxuries when it judges that a user is stressed or feels like a failure, based upon the language of their posts, messages and click history.
3. A recommender system that presents users with posts that many would find offensive, in order to determine whether the user would like to see that content (like a bandit algorithm), but does not show them again unless the user engages with them.

1. A recommender system that presents users with offensive posts based upon its estimate of the gender, race, sexuality, etc of the user.

The Ethics of Recommender Systems

Example	Average value on a scale of 1 (no ethical problems) to 5 (highly unethical)
A recommender algorithm that presents users with advertisements similar to the posts they have clicked on in the past.	
A recommender system that chooses advertisements based on whether it judges that a user is stressed or feels like a failure.	
A recommender system that presents users with shocking content in order to determine whether the user would like to see that content (like a bandit algorithm).	
A recommender system that presents users pictures of cute animals over and over again, to the exclusion of other content, when the user has a compulsion to look at pictures of cute animals.	

Issue 2: The Feed

- One concept that might help make sense of your intuitions on the previous questions is **manipulation**.

Example 1: Conditioning

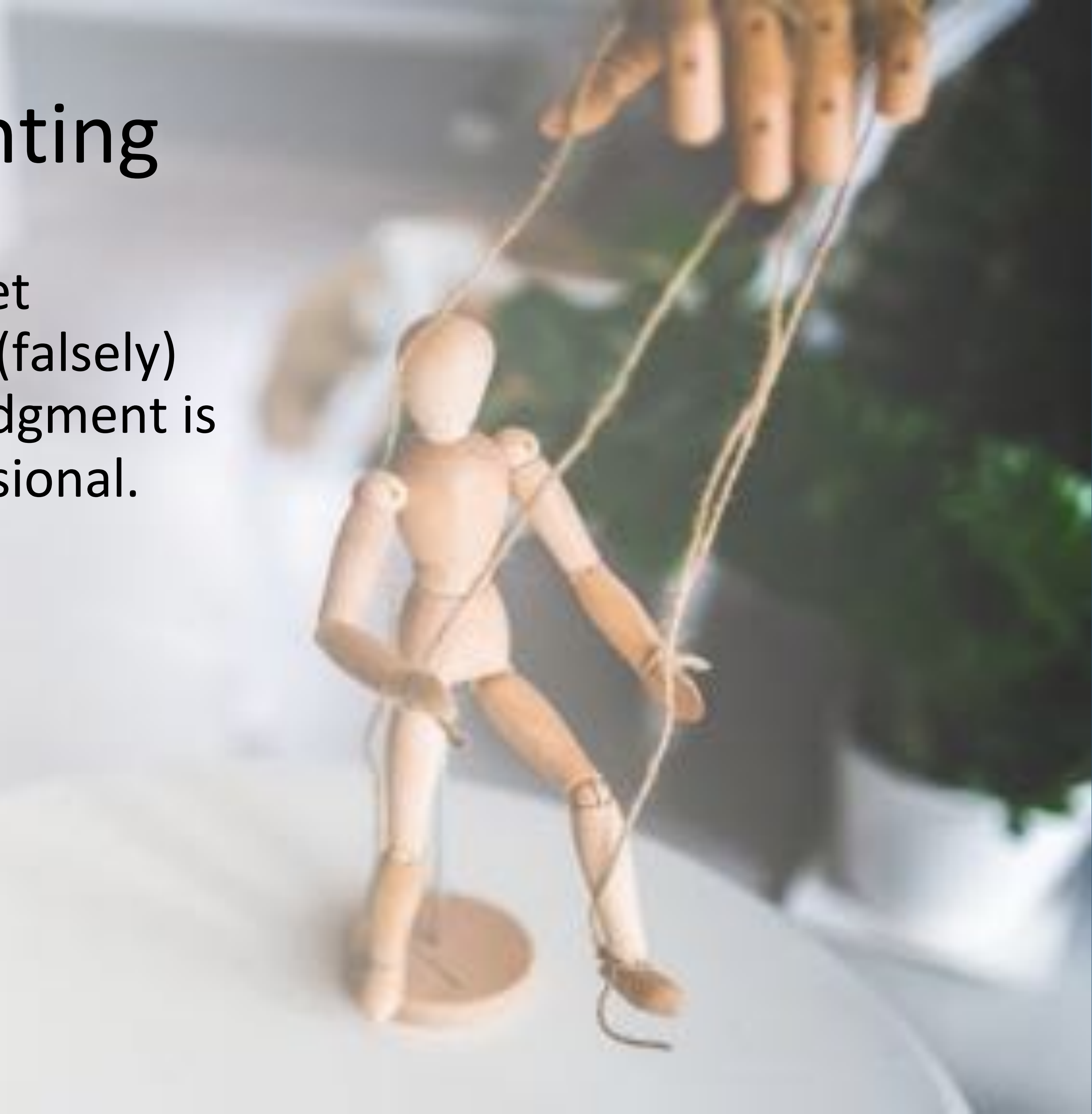
- Conditioning is an attempt to get someone to adopt a pattern of behaviour by rewarding or punishing their actions.

Example 2: The Guilt Trip

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

Example 3: Gaslighting

- Gaslighting is an attempt to get someone to do something by (falsely) persuading them that their judgment is generally flawed or even delusional.



Defining Manipulation

- A **definition** of manipulation would explain what all of these cases have in common with each other.
- **In the chat**: What do the previous three examples have in common with each other that make them count as 'manipulation'?

Defining 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”)

Defining Manipulation

- What norms or ideals guide our beliefs, desires and emotions? Noggle:
- Beliefs:

“Believe only the truth.”

- Desires:

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

- Emotions:

“Base your emotions on true beliefs.”

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

Defining Manipulation

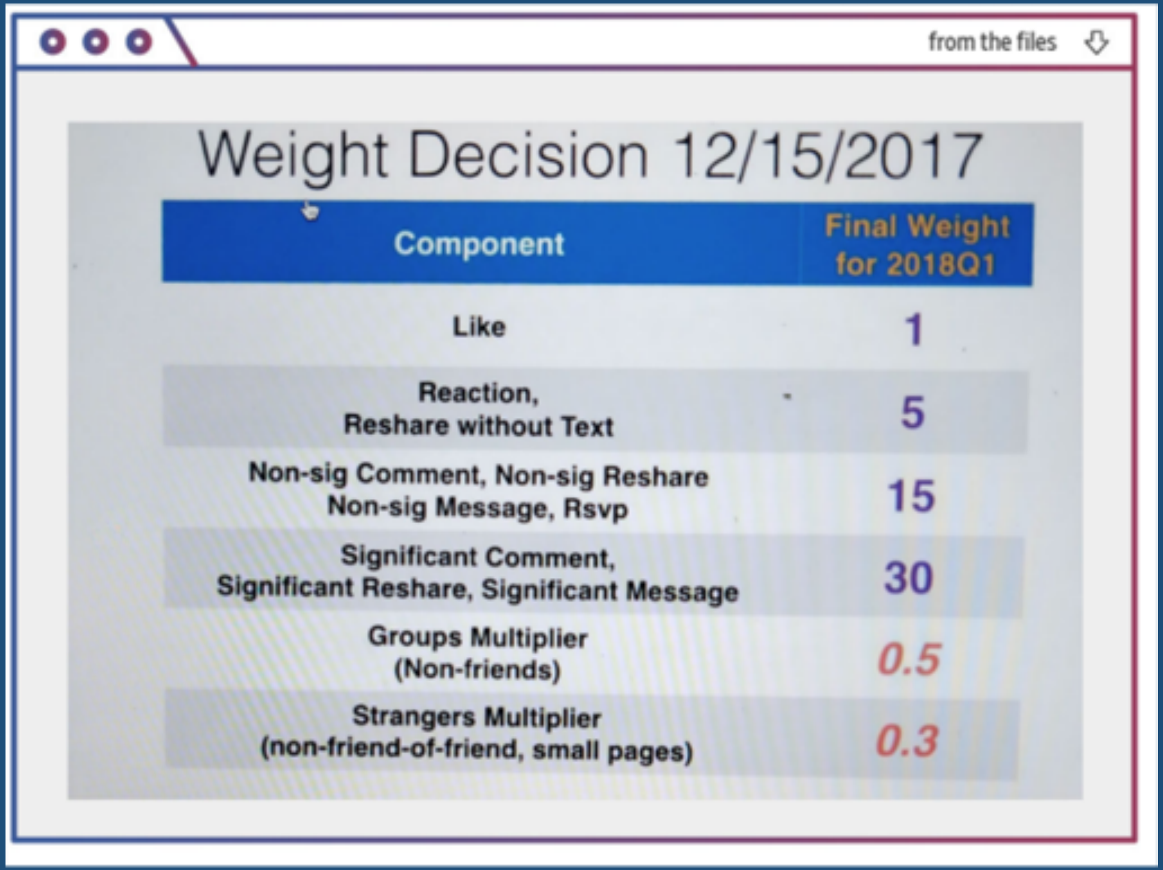
Example	Does it violate the norms of our beliefs, desires, emotions, or none of them?
Conditioning	
The guilt trip	
Gaslighting	
A recommender system that chooses advertisements based on whether it judges that a user is stressed or feels like a failure.	
A recommender system that presents users with shocking content in order to determine whether the user would like to see that content (like a bandit algorithm).	
A recommender system that presents users pictures of cute animals over and over again, when the user has a compulsion to look at pictures of cute animals, to the exclusion of other content.	

The Morality of Influencing Others

Three plausible options for the ethics of manipulation:

1. Intentional manipulation is always unethical.
2. Intentional manipulation is “prima facie” or “pro tanto” unethical.
   - Manipulation is always unethical unless it is required to fulfill another moral duty
3. Intentional manipulation is not inherently unethical; it is only bad when it leads to bad outcomes.

Back to the Second Issue

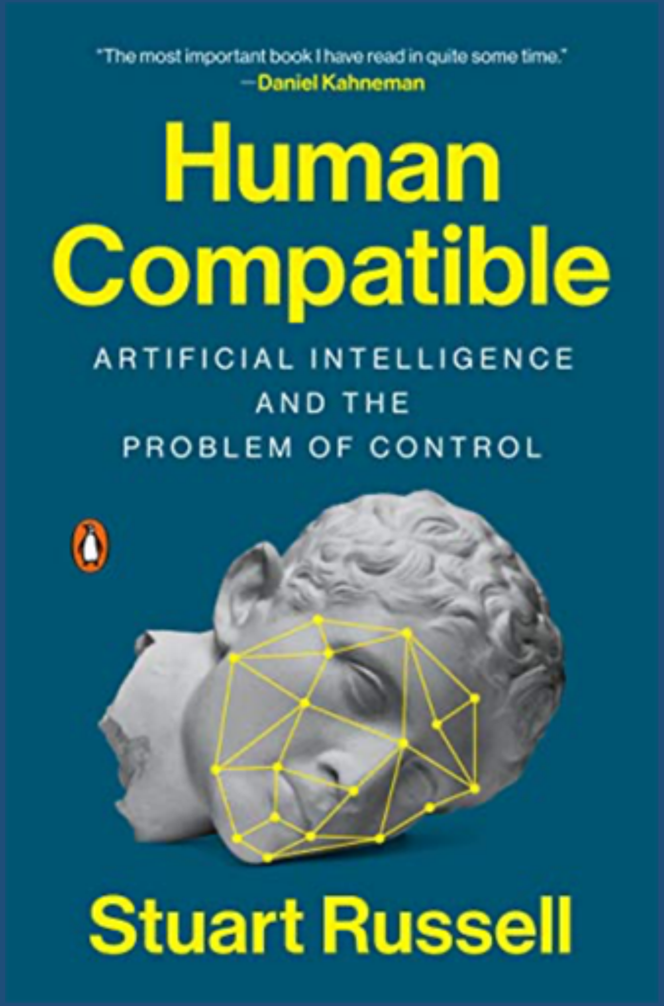


Weight Decision 12/15/2017

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*

The concepts discussed so far may help us to better understand our reactions to the second issue revealed by the Facebook Papers: the amplification of angry posts. Some possible reactions:

1. The amplification of angry content is ethical (or at least 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 for some other reason.


"The most important book I have read in quite some time."
—Daniel Kahneman
Human Compatible
ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL
Stuart Russell

Stuart Russell, the writer of the best-known AI textbook, has hypothesized the following:

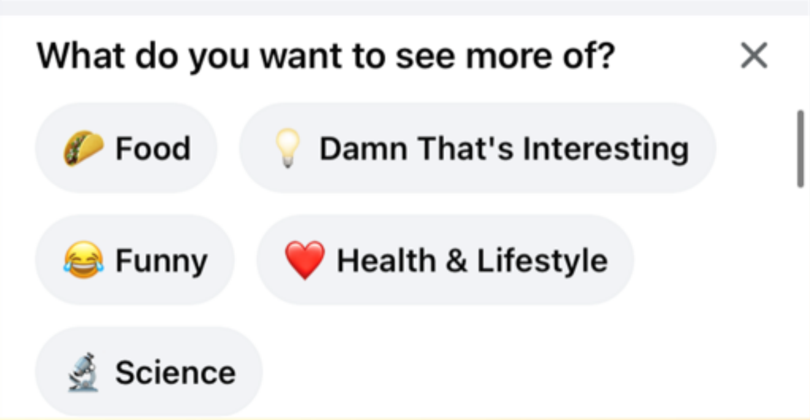
"Typically, such algorithms are designed to maximize click-through, that is, the probability that the user clicks on presented items. The solution is simply to present items that the user likes to click on, right? Wrong. The solution is to change the user's preferences so that they become more predictable. A more predictable user can be fed items that they are likely to click on, thereby generating more revenue. People with more extreme political views tend to be more predictable in which items they will click on…. Like any rational entity, the algorithm learns how to modify the state of its environment – in this case, the user's mind – in order to maximize its own reward." (Russell, *Human Compatible*, 8)

Improving Recommender Systems?

- **Consent** may lead to more ethical recommender systems.
- If a user consents to be influenced in a certain way, perhaps that makes it more permissible, even if that influence involves manipulation.
- Legal requirements vs moral requirements
- Explicit consent > implicit consent

Improving Recommender Systems?

- **Giving users control** over what they see in their feeds may also lead to more ethical recommender systems. (Stray, ‘Beyond Engagement’)
- E.g. ‘see less often’ or ‘hide post’ functions in feeds



What do you want to see more of?

Discussion Question

In the chat (or raise your hand to speak):

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

Improving Recommender Systems?

- The most radical change: we might design more ethical recommender systems by changing the **objective function** of a recommender system – what it is that the recommender system is trying to maximize. (Stray, ‘Beyond Engagement’)
- Well-being metrics
- IEEE 7010

Other Resources

- Talks and events at:
  - Centre for Ethics
  - Schwartz Reisman Institute
- Courses in:
  - Philosophy
  - History and Philosophy of Science and Technology
  - Ethics, Society and Law
  - Faculty of Information
- CSC 300: Computer Science and Society

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.

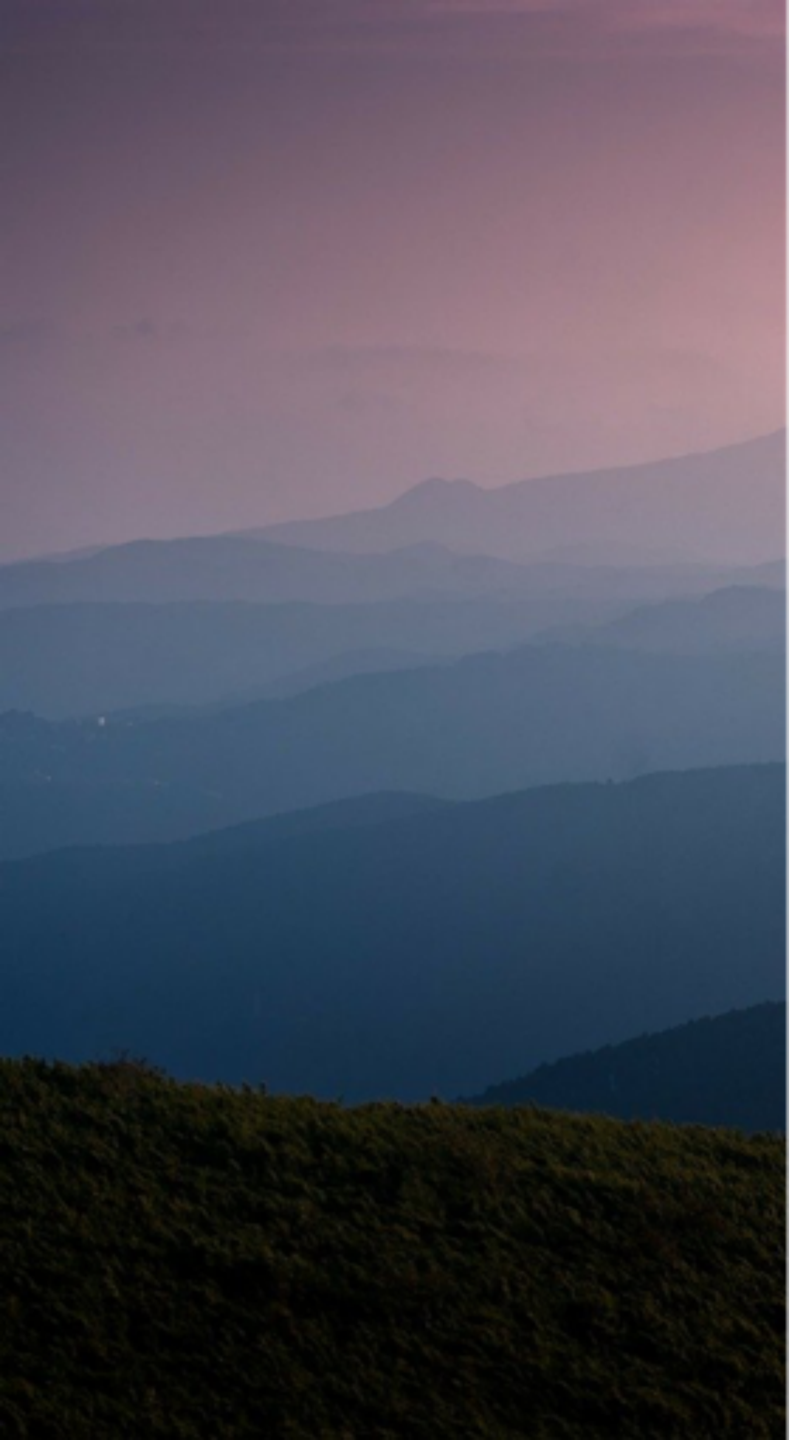
Instructional Team:

Roger Grosse, Steven Coyne, Emma McClure

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



References

- Mill, John Stuart. *On Liberty.* Accessed online: https://socialsciences.mcmaster.ca/econ/ugcm/3ll3/mill/liberty.pdf
- 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
- Jonathan Stray, “Beyond Engagement” Accessed online: https://partnershiponai.org/beyond-engagement-aligning-algorithmic-recommendations-with-prosocial-goals/