Content Moderation

Module 6 of a course on *Ethical Issues in AI*

Prepared by

John Hooker
Emeritus Professor, Carnegie Mellon University

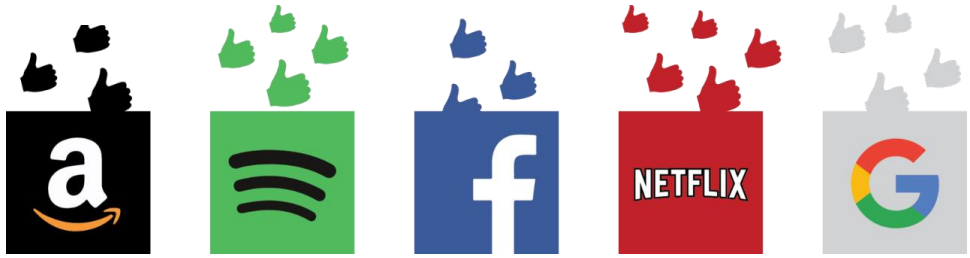
CMU Osher, January 2025

AI recommender systems

- Recommender systems are the **chief means of content moderation**.
  - *They decide what you see.*
  - *Social media recommenders (“algorithms”)*
  - *News media recommenders (“click bait”)*
  - *Retails sites, search engines, GPTs*

AI recommender systems

- One of the most **effective** and **widely-adopted** AI applications
  - *Many successful techniques*
    - Matrix factorization
    - Bayesian classifiers and decision trees
    - Collaborative filtering, *k* nearest neighbors
    - Recurrent neural networks and transformers, etc.
  - *A **major force** in business, marketing, and news media*



AI recommender systems

- AI recommenders are trained to **maximize engagement**.
  - *The result is the worldwide spread of*
    - Lies, slander, hate speech, harmful misinformation
    - Sensational news coverage
- Calls for **content moderation**
  - *To avoid harmful side effects of profit-maximizing content algorithms.*

Factoid: **Lies** spread 6.5 times faster on social media than **truth**.

AI recommender systems

- Forms of content moderation:
  - ***Taking down*** *material.*
  - ***Putting material at the end*** *of the recommended list.*
    - This, in effect, is the same as taking it down
  - ***Flagging*** *material as false or offensive*.
- Recommender systems can implement all of these.

Factoid: **Lies** spread 6.5 times faster on social media than **truth**.

Content moderation

- Current policies are **ad hoc**
  - *Respond to complaints.*
  - *Change with ownership and political climate.*
  - *We need a **principled approach**.*

Content moderation

- Current policies are **ad hoc**
  - *Respond to complaints.*
  - *Change with ownership and political climate.*
  - *We need a **principled approach**.*
- Rather than try to resolve all the issues, we focus on two **case studies**:
  - *Inciting violence on YouTube*
  - *Social media impact on young people.*

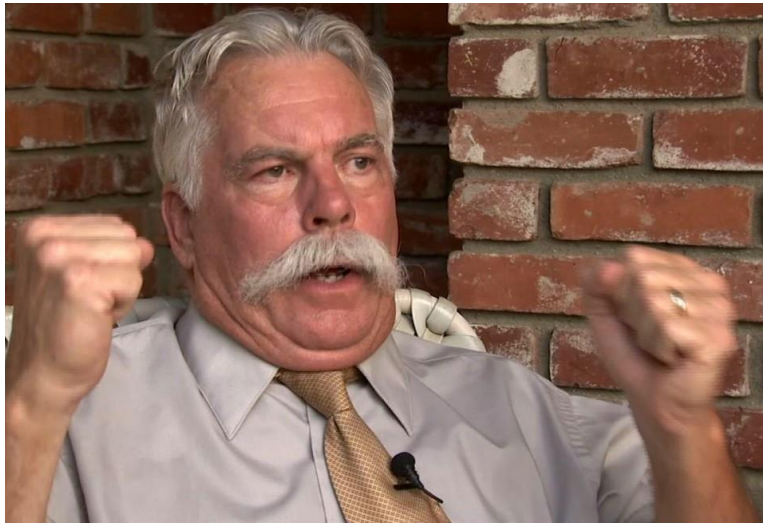
Content moderation

- Rather than try to resolve all the issues, we focus on two **case studies**:
  - *Inciting violence on YouTube*
  - *Social media impact on young people.*
- Note:
  - *We are not talking about government regulation.*
  - *Only about how online platforms should **voluntarily** regulate content.*

Case study: Inciting violence

Case study: Inciting violence

- A very high-profile dilemma
  - *Raised issues that have not been resolved* ***to this day****.*
  - *An amateurish film,* Innocence of Muslims*, was uploaded to* ***YouTube*** *on 1 July 2012.*
    - Packed with **lies** and misinformation.
    - Highly **offensive** due to negative portrayal of Islam.



Case study: Inciting violence

- Reaction...
  - *Protests worldwide, some **violent**.*
  - *Reportedly 50 deaths, mainly in Pakistan.*
  - *Most protests anti-U.S. because film maker lived in the U.S.*
  - *U.S. government didn't ban the video.*

Case study: Inciting violence

- President Barack Obama asked Google (owner of YouTube) to take down the film.
  - *But he had no legal authority to require it.*
  - *Google **refused**.*
    - …but removed the video in some countries.

Case study: Inciting violence

- Google's position:
  - *The post is consistent with "company policy."*
  - *"It is against the Islam religion [sic] but not Muslim people."*
    - As reported in NY Times, 14 Sep 2012.
    - YouTube now has a laundry list of "community guidelines," but what is the principle?

Case study: Inciting violence

- **Generalizability**
  - *Is **allowing lies to be posted** a form of **deception**?*
    - Does the **mere fact** that the lies are posted **cause** people to believe something YouTube knows is false?
    - Only if users regard the mere appearance of the video as a **claim** or **endorsement** by YouTube.
    - But people know that YouTube allows **all sorts** of **contradictory** views to be posted.
    - So, there is no obvious endorsement.
    - It is hard to make a case that **YouTube** (as opposed to those who upload the lies) is deceiving people.

Case study: Inciting violence

- **Generalizability**
  - *Is YouTube's **rationale** for allowing lies to be posted generalizable? Perhaps.*
    - Would YouTube continue to **achieve its purposes** even if all online sites allowed lies to be posted?
    - Probably. Their purpose is to **make money**, not to convince anyone to believe their content.
    - Practically all online sites **already** allow lies to be posted, and YouTube continues to make tons of money.

Case study: Inciting violence

- **Generalizability**
  - *Is **removing** videos generalizable? Depends on the **reason**.*
    - Removing **false videos because they reduce utility** may not be generalizable.
    - Generalized private censorship of **information** may destroy more utility than it creates.

Case study: Inciting violence

- **Generalizability**
  - *Is **removing** videos generalizable? Depends on the **reason**.*
    - Removing **false videos because they reduce utility** may not be generalizable.
    - Generalized private censorship of **information** may destroy more utility than it creates.
    - But removing **incendiary** videos because they **may incite violence** is generalizable.
    - Removal of an incendiary video would **still** reduce probability of violence even if it were general practice to do so.

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument 1* ***against*** *content moderation.*
    - **YouTube** didn't kill anyone in a riot.
    - The rioters did.

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument 1* ***against*** *content moderation.*
    - **YouTube** didn't kill anyone in a riot.
    - The rioters did.
  - *Response.*
    - For the utilitarian principle, only the **consequences** of YouTube's policy matter.
    - It doesn't matter whether the consequences involve the conduct of other people.

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument 2* ***against*** *content moderation.*
    - "Ought implies can."
    - Content moderation is a **massive task**.
    - YouTube: **500 hours** of video uploaded **per minute**.
    - Facebook: **3 billion** people log in **every day**.
    - It is **impossible** to screen so much content.
    - The utilitarian principle only requires one to consider **available actions**.

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument 2 – Response.*
    - This is a **factual** claim, not an ethical one.
    - Anyway, online sites **already use AI-based** content moderation.
    - This is how they direct **relevant content** to users!
    - AI can **flag** questionable posts for human moderation.
    - **Users** also request takedowns.*

**Although mainly for alleged copyright infringement.*

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument 2 – Response.*
    - This is a **factual** claim, not an ethical one.
    - Anyway, online sites **already use AI-based** content moderation.
    - This is how they direct **relevant content** to users!
    - AI can **flag** questionable posts for human moderation.
    - **Users** also request takedowns.*
    - Social media companies already employ thousands of **content moderators** and can afford more.
    - YouTube: **10,000** content moderators.
    - Facebook: **15,000** content moderators, out of **180,000+** employees total (mostly in ad sales & revenue generation).
    - About **100,000** content moderators worldwide

**Although mainly for alleged copyright infringement.*

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument **for** content moderation*
    - It **increases utility**.
    - Even if it's impossible to catch **all** harmful posts, it's possible to catch **many** of them.
    - Failure to do so is **clear violation** of utilitarian principle.

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument **for** content moderation*
    - It **increases utility**.
    - Even if it's impossible to catch **all** harmful posts, it's possible to catch **many** of them.
    - Failure to do so is **clear violation** of utilitarian principle.
  - *Response – We are **already doing all we can**.*
    - For example, we took down all Covid vaccine misinformation.
    - If so, great. Then you agree that you **should** do so?
    - Anyway, you can clearly use **existing recommender technology for different ends**.

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument 3 **against** content moderation – Free speech.*
    - Content moderation violates the **First Amendment rights** of users.
    - This is a **generalization** argument. Anyway, the First Amendment of the U.S. Constitution forbids **government** from restricting free speech.
    - YouTube is a **private company**.

Case study: Inciting violence

- **Utilitarian** analysis
  - *Argument 3 **against** content moderation – Free speech.*
    - Content moderation violates the **First Amendment rights** of users.
    - This is a **generalization** argument. Anyway, the First Amendment of the U.S. Constitution forbids **government** from restricting free speech.
    - YouTube is a **private company**.
  - *Revised argument 3*
    - Content moderation **restricts free speech**, and this is harmful to society.
    - Is it a restriction, or a refusal to give one a particular megaphone?

Case study: Inciting violence

- **Autonomy**
  - *This is the most demanding principle.*
    - Malicious rumors, terrorist posts, medical misinformation, etc., can lead to **death and injury** – e.g., in riots.
    - Isn't this a **violation of autonomy**?

Case study: Inciting violence

- **Autonomy**
  - *This is the most demanding principle.*
    - Malicious rumors, terrorist posts, medical misinformation, etc., can lead to **death and injury** – e.g., in riots.
    - Isn't this a **violation of autonomy**?
  - *Common response*.
    - **YouTube** didn't kill anyone in a riot. The rioters did.
    - This is not enough to escape the autonomy principle…

Case study: Inciting violence

- **Autonomy**
  - *Let's apply the principle:*
    - A post should be removed when one is **rationally constrained to believe** debilitating harm will result.
    - Regardless of who immediately causes the harm.

Case study: Inciting violence

- **Autonomy**
  - *Let's apply the principle:*
    - A post should be removed when one is **rationally constrained to believe** debilitating harm will result.
    - Regardless of who immediately causes the harm.
  - *Response:*
    - The rioters gave **implied consent**: they voluntarily assumed the risk of joining a riot.
    - So, there is **no autonomy violation**.
    - But… were **innocent bystanders** hurt?
    - If so, we have a violation.

Case study: Inciting violence

- **Autonomy**
  - *The autonomy principle can be even more demanding…*
    - Perhaps YouTube managers are rationally constrained to believe that the **very existence** of YouTube will, at some point, interfere with ethical action plans without implied consent, **despite their best efforts** to remove offensive videos.
    - If so, the site should be **shut down**.
    - To stay in business ethically, YouTube managers must be so thorough in their content moderation that it is **not irrational for them to believe** that this will **never happen**.

Case study: Inciting violence

- **Take down misinformation in general?**
  - *Not required by autonomy principle*
  - *Could be utilitarian, but…*
  - *Could also be ungeneralizable.*
  - *Option: **flag** what fact checkers see as misinformation*
    - …while providing reliable sources.
    - Does not stifle free speech, and may promote it.
    - May be **required** by utilitarian principle

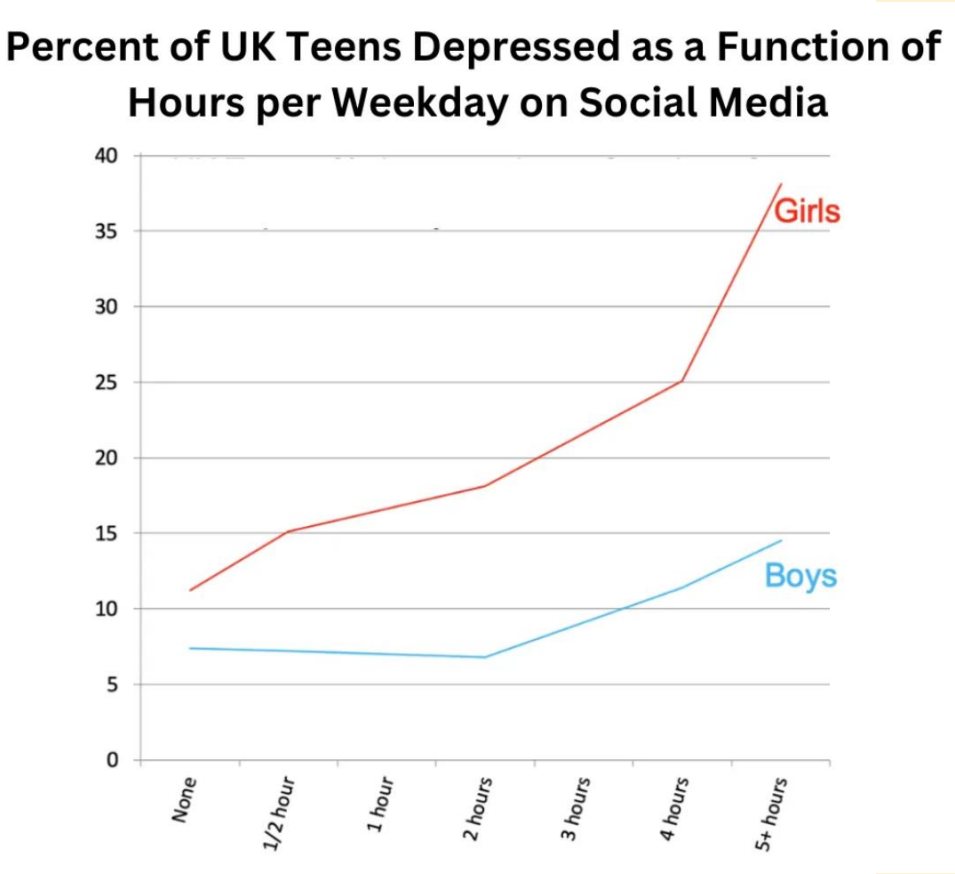
Case study: Inciting violence

- **To sum up…**
  - *YouTube must **shut down** unless it adopts a content moderation policy for which YouTube managers can rationally believe that **autonomy violations** are **not inevitable**.*
  - *There is a **strong utilitarian imperative** to identify such a policy, due to the many benefits of YouTube.*
  - *A policy of removing **false claims** must be carefully crafted to avoid violating **generalizability**.*
  - *A compromise is to **flag** clearly false claims, while providing reliable sources.*

Case study: Impact of social media on young people

Case study: Impact on young people

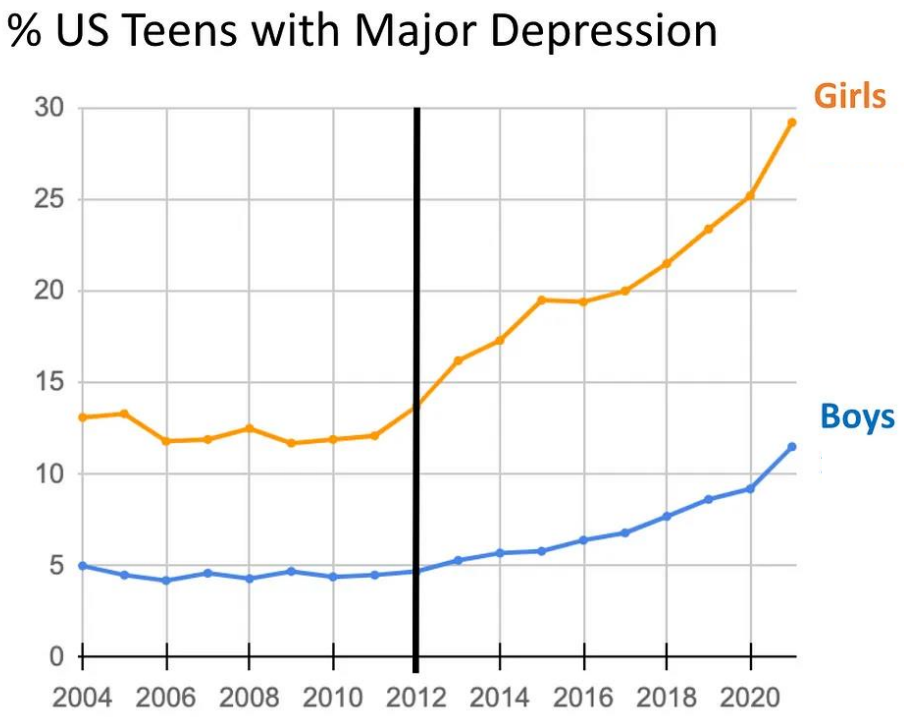
- Depression, anxiety and suicide.
  - *All are rising among young people in some countries*
  - *This roughly coincides with rise of social media use.*



Source: Data in Table 2, Y. Kelley, A. Zilanawala, C. Booker, A. Sacker, Social media use and adolescent mental health: Fimdings from the UK Millenium cohort Study, *The Lancet*, 2018.

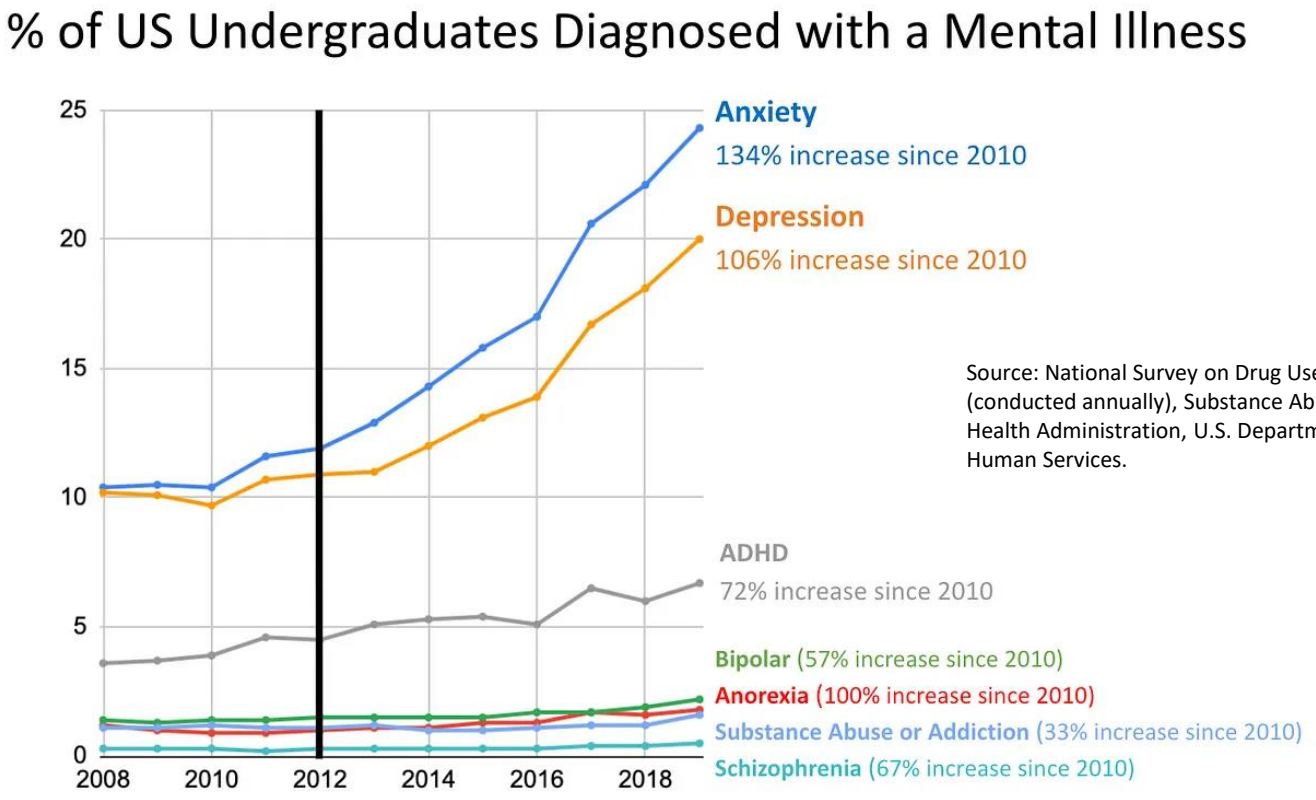
Case study: Impact on young people

- Depression, anxiety and suicide.
  - *All are rising among young people is some countries*
  - *This roughly coincides with rise of social media use.*



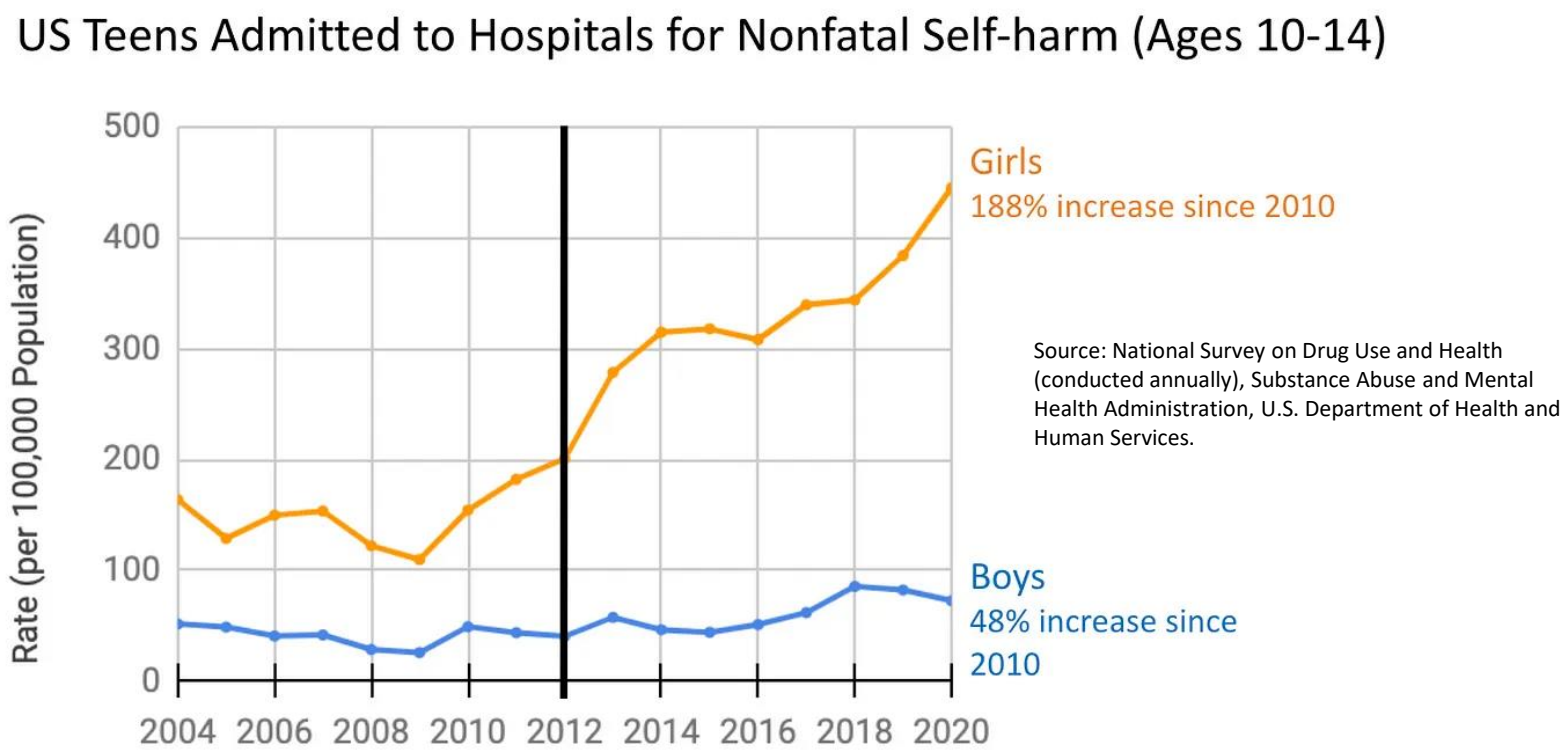
Source: National Survey on Drug Use and Health (conducted annually), Substance Abuse and Mental Health Administration, U.S. Department of Health and Human Services.

Case study: Impact on young people



Source: National Survey on Drug Use and Health (conducted annually), Substance Abuse and Mental Health Administration, U.S. Department of Health and Human Services.

Case study: Impact on young people



Case study: Impact on young people

- *These data have other interpretations.*
  - Heavy social media use is the **result** of depression rather than the cause.
  - Rising depression is due to more frequent **reporting** and acknowledgment of psychological problems.
  - Rising depression is real but has **other causes** that happen to coincide with social media use.
- *We can't resolve the factual issues.*
  - But we can ask: **Assuming** that social media overuse causes depression, what should companies do about it?

Case study: Impact on young people

- Similar to previous case study, but with a twist:
  - *It concerns **young** users, generally minors for legal purposes.*
  - *Let's think more carefully about autonomy issues…*

Case study: Impact on young people

- Autonomy issues and children.
  - *First, denying access to YouTube (whether child or adult) is **no violation of autonomy**.*
    - Users cannot have an **action plan** of being granted access to YouTube.
    - **Only YouTube** can decide whether to grant access.
    - The user can only decide to access YouTube **if** access is granted.

Case study: Impact on young people

- Autonomy issues and children.
  - *Second, children are not **fully autonomous**.*
    - They often do not (or cannot) form a **coherent rationale** for their behavior.
    - In such cases, parents can **forbid** the behavior without violation of autonomy.

Case study: Impact on young people

- Autonomy issues and children.
  - *Third, children are nonetheless **agents**.*
    - They **sometimes** act autonomously.
    - So, injuring a child **violates autonomy** (as well as the utilitarian principle)

Case study: Impact on young people

- Autonomy issues and children.
  - *Third, children are nonetheless **agents**.*
    - They **sometimes** act autonomously.
    - So, injuring a child **violates autonomy** (as well as the utilitarian principle)
  - *This is important because...*
    - Utilitarian benefits of allowing children online can **never outweigh** autonomy violations.
    - Children are much less capable than adults of giving **informed consent** to the risk of injury.
    - So, there is a **greater chance** of autonomy violations.

Conclusions…

- Generalization principle
  - ***Removing*** *clearly harmful content is* ***generalizable****.*
  - ***Failure*** *to remove it is* ***also generalizable****.*
  - *A carefully crafted policy of removing* ***lies*** *and other clear* ***falsehoods*** *may be generalizable.*
  - ***Flagging*** *lies is generalizable.*
- Autonomy principle
  - *Online sites can* ***ethically operate*** *only if one can rationally believe that their moderated content* ***will never violate autonomy*** *without informed consent.*
  - *Since* ***children*** *are agents, their* ***autonomy must be protected****.*

Conclusions…

- Utilitarian principle
  - *A site should make a **concerted effort** to find an ethical content moderation policy, rather than shut down.*
  - *It should at least **flag** false and misleading content, while citing reliable sources.*