Panel 8: Vulnerability and Control / Vulnérabilité et contrôle

Deepfakes, Image Based Abuse, and Online Harm
Panelists

• Eve Gaumond (Law Student, Faculté de droit, Université Laval, @eve_gaumond)
• Suzie Dunn (PhD Candidate & Part-Time Professor, Faculty of Law, University of Ottawa @SuzieMDunn)
• Nareg Froundjian (Lawyer, Technology Law, @naregeff)
• Yuan Stevens (Research Consultant, Faculté de droit, Université de Montréal and Data & Society Research Institute, @ystvns)
Positive impacts of online interaction? 😊
Audience question: What kind of online-harm have you or others around you experienced in your professional or personal life?
Audience question: If you spot fake news on social media, how do you react?

- 75% Ignore
- 6% Report
- 8% Debunk
- 12% Rant to friends
70,971
Google searches
76,481
YouTube videos viewed
8,282 Tweets sent
882

Instagram photos uploaded
1. **Moderation activity**: constitutionally inherent to social platforms
2. **Normativity on social media**: complex, in conjunction rather than silos
3. **Easy or daunting**: self-censorship vs free and open space
4. **Liability regimes**: balancing of individual harms and public interest
1. Moderation Activity
The need to moderate

User expectations
- no harassment
- no racism
- no graphic violence
- no obscenity
- no copyrighted material
- no viruses
- no disinformation

Platform promises
- open, inviting spaces
- for public dialogue
- content agnostic
- bastions of free speech
Reasons

- Nudity or sexual content
- Harmful or dangerous content
- Hateful content
- Violent or graphic content
- Harassment and cyberbullying
- Spam, misleading metadata, scams
- Threats
- Copyright
- Privacy
- Impersonation
- Child safety
- More...
2. Normativity on Social Media
Top-down Community Guidelines

“Here are some common-sense rules that'll help you steer clear of trouble. Please take these rules seriously and take them to heart. Don't try to look for loopholes or try to lawyer your way around the guidelines—just understand them and try to respect the spirit in which they were created.”

[youtube guidelines]
Crowdsourced normativity on reddit and /r/

Reddiquette is an informal expression of the values of many redditors, as written by redditors themselves. Please abide by it the best you can.

/r/AskHistorians aims to provide serious, academic-level answers to questions about history.

We have written these rules to support this aim and maintain the high standard of discussion this subreddit has become known for.

Please note that /r/AskHistorians is actively moderated. Moderators regularly take action to enforce these rules.
Hello everyone,

In this thread, there have been a large number of incorrect, speculative, and often unhelpful responses. Many are asking about the deleted comments, which merely compound the problem. Please, before you attempt to answer the question, keep in mind that we are not looking for comprehensive answers that break the rules. This thread is intended to be a place where you can ask questions and share your thoughts, but it can be frustrating to come in here from your front page and see some inappropriate posts.

This reply is not appropriate for this subreddit. While we aren't as humorless as our reputation implies, a comment should not consist solely of a joke, although incorporating humor into a proper answer is acceptable. Do not post in this manner again.

In addition, questions or concerns, I would ask that they be directed to modmail, or a META thread. Thank you!
WARNING:
Some of the following images are graphic in nature and might be disturbing to some viewers.
Signs at Facebook Headquarters

- Reduce Clickbait
- Depolarize
- Reduce Misinfo
- Unship Hate

Image by Jason Koebler
This video is no longer available due to a copyright claim by Lagpreacher. Sorry about that.
Free expression is paramount, but there are times when speech can be at odds with authenticity, safety, privacy, and dignity. Some expression can endanger other people’s ability to express themselves freely.

Therefore, it must be balanced against these considerations. In light of this balance, internet services have a responsibility to set standards for what is and is not acceptable to share on their platforms.

Those standards should protect people and their expression, and any limits should be based on specific values that companies have the responsibility to articulate.

To ensure fair decision-making based on standards and values, internet services can establish bodies designed to oversee important matters of expression and to make independent final decisions.
3. Easy or Daunting?
Values are easier to define in small groups (Ivy League!) – are there global values?

The Facebook Wall

What does community mean to you?

Community means staying connected.
Is it nudity? Is it art? Where do you draw your line?
Is there such a thing as clearly offensive content?

Allen Henning  
(British)
whatilmyusernamebe 10:32 PM
Hey man, I think you're a generally well-intentioned dude, but why do you admins not just ban hate speech? There's no reason not to. Seriously. These people can't be reasoned with. You're not protecting free speech, but you are making Reddit look absolutely awful in the media. Tell them to go somewhere else, or you and everybody that works at Reddit is officially endorsing the hate speech they allow.

spez 10:41 PM
Hate speech is difficult to define. There's a reason why it's not really done. Additionally, we are not the thought police. It's not the role of a private company to decide what people can and cannot say.

whatilmyusernamebe 10:42 PM
I'd argue that hate speech should be banned with its own rule, separate from the violence policy.

spez 10:44 PM
But it *is* the role of a private company to decide what people can and cannot say *on their platform*.

I know what you're asking, but it's a nearly impossible precedent to uphold. It's impossible to enforce consistently.
In a global world, it is less about being right or fair and more about being transparent and consistent.
4. Liability Regimes
(1) Treatment of publisher or speaker
No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.

(2) Civil liability
No provider or user of an interactive computer service shall be held liable on account of—

(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or

(B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).

[s. 230 of the Communication Decency Act]
22. Le prestataire de services qui agit à titre d'intermédiaire pour offrir des services de conservation de documents technologiques sur un réseau de communication n'est pas responsable des activités accomplies par l'utilisateur du service au moyen des documents remisés par ce dernier ou à la demande de celui-ci.

Cependant, il peut engager sa responsabilité, notamment s'il a de fait connaissance que les documents conservés servent à la réalisation d'une activité à caractère illicite ou s'il a connaissance de circonstances qui la rendent apparente et qu'il n'agit pas promptement pour rendre l'accès aux documents impossible ou pour autrement empêcher la poursuite de cette activité.

[...]

Loi concernant le cadre juridique des technologies de l'information, RLRQ, c-1.1
Art. 18. Internet connection providers do not have civil liability for damages resulting from content produced by third parties.

Art. 19. In order to ensure freedom of expression and prevent censorship, Internet applications providers may only be held civilly liable for damages resulting from content generated by third parties if, after specific judicial order, the provider fails to take action to make the content identified as offensive unavailable on its service by the stipulated deadline, subject to the technical limitations of its service and any legal provisions to the contrary.
Art. 20. If the Internet application provider has contact information for the user who is directly responsible for the content referred to in article 19, the provider must notify the user for the reasons for removing the content and other information related to its removal, with sufficient detail to enable a full answer and defense in court, unless applicable legislation or a reasoned court order expressly stipulates otherwise.
Art. 20. §1. At the request of the user who posted the content that was removed, the Internet applications provider, if it is a legal entity providing applications in an organized, professional manner, for profit, must replace the removed content with a statement of the reasons for removal or the judicial order to remove the content.
Art. 21. Internet applications providers that make available content created by third parties will be secondarily liable for the violation of privacy resulting from the disclosure, without the participants’ authorization, of images, videos, and other materials containing nudity or sexual acts of a private nature, if after receiving notice from the participant or the participant’s legal representative, the Internet applications provider fails to promptly to remove the content from its service, subject to technical limitations of the service.
a brief (hi)story

@ystvns
“non-consensual distribution of intimate images”
gender-based violence online; “doxing”
the impact
‘It’s an abuse of me and my body. It feels like it’s sexual abuse’
Lucy

‘It’s still a picture of you ... it’s still abuse’
Stakeholder working with victim-survivors

‘It’s just this panic that something is going to happen ... I think like the second that I’m not prepared for it, then it’s going to happen’
Stephen

‘It’s a type of rape, it’s just the digital version’
Deborah

the law
- s. 162.1 (criminal code)
- various **provincial** laws re: intimate images
- **copyright** law
- **class** actions
- breach of **contract**

More than “Revenge Porn”:
*Civil Remedies for the Non-consensual Distribution of Intimate Images, S. Dunn & A. Petricone-Westwood (2018)*
• tort law (appropriation of likeness, breach of confidence, breach of fiduciary duty, defamation, extortion/intimidation, harassment, intentional infliction of mental suffering, intrusion upon seclusion, publication of private facts)

online = irl
vulnerable / vulnerabilities
institutional bypass?
“Instead of trying to fix dysfunctional institutions, as most failed reforms do, they simply bypass them. ... Like a “coronary bypass” surgery, an institutional bypass creates new pathways around clogged or blocked institutions.”

Institutional Bypass: An Alternative for Development Reform, M. Mota Prado (2011)
Founded in August of 2017, BADASS is a nonprofit organization dedicated to providing support to victims of revenge porn/image abuse, and eradicating the practice through education, advocacy, and legislation. Our goal is to arm victims with the tools they need to become their own advocates for justice, and provide the resources they need to regain control of their images, empower themselves, and get justice.
Detecting Non-Consensual Intimate Images and Supporting Victims
New federal investment will help end cyberviolence

From: Women and Gender Equality Canada

News release

Project will empower women and girls to assess threats and promote responsible digital citizenship

August 27, 2019 – Ottawa, Ontario – Women and Gender Equality Canada

Canada’s largest science, technology, engineering and mathematics (STEM) outreach organization, Actua, will receive up to $600,000 to reduce cyberviolence and promote responsible digital and community citizenship. Actua will develop innovative on- and off-line programming that empowers girls and young women to critically assess online interactions and threats in order to reduce cyberviolence and promote responsible digital and community citizenship.
Who is most targeted by deepfake videos?

- Politicians (Male): 16
- Celebrities (Male): 1
- Everyday people (Male): 0
- Politicians (Female): 7
- Celebrities (Female): 16
- Everyday people (Female): 6
DEEPTRACE AND THE STATE OF DEEPFAKES

Gender
Deepfake pornography is a phenomenon that exclusively targets and harms women. In contrast, the non-pornographic deepfake videos we analyzed on YouTube contained a majority of male subjects.

Nationality
We found that over 90% of deepfake videos on YouTube featured Western subjects. However, non-Western subjects featured in almost a third of videos on deepfake pornography websites, with South Korean K-pop singers making up a quarter of the subjects targeted. This indicates that deepfake pornography is an increasingly global phenomenon.

Profession
All but 1% of the subjects featured in deepfake pornography videos were actresses and musicians working in the entertainment sector. However, subjects featuring in YouTube deepfake videos came from a more diverse range of professions, notably including politicians and corporate figures.
SYNTHETIC MEDIA
DEEP FAKE
FACIAL RE-ENACTMENT
AUDIO GENERATION
ORIGINS

FILM INDUSTRY CGI
REDDIT FAKEAPP
SCHOLARS ARMS RACE
TECH COMPANIES DETECTION
THANK YOU!!!

Eve Gaumont (Law Student, Faculté de droit, Université Laval, @eve_gaumont,)

Suzie Dunn (PhD Candidate & Part-Time Professor, Faculty of Law, University of Ottawa @SuzieMDunn)

Nareg Froundjian (Lawyer, Technology Law, @naregeff)

Yuan Stevens (Research Consultant, Faculté de droit, Université de Montréal and Data & Society Research Institute, @ystvns)
Any questions for the panelists?

- Substantive law reform is needed but what about remedies and the whack-a-mole problem?
- What advice do you have for responsible sharing of intimate images?
- Do you see a new role for the United Nations to facilitate a dialogue around acceptable uses of social media which could lead to good laws?
- Croyez-vous que le droit est le meilleur outil pour répondre à ces nouveaux enjeux?
- How should these technologies influence the development of evidence law?
- How does the sharing of non consensual images and deepfakes play in employment situations?
- Par rapport à la perte de confiance dont parlait Me Dunn dans sa conclusion, pensez-vous que l’on devrait expressément prohiber les deepfakes dans la loi?
- Should algorithms be audited for being legal / do no harm or regulated to be applied for legal purposes?
- Do you believe that current paradigms of law, based largely on territorial silos rather than on a networked environment, are adequate to ensure a meaningful regulation of these practices?