# Platform Governance: Regulating Social Media (615i513a)

Dr. Natalia Umansky

2022-2023

---

**Time:** Monday 16:15 - 18:00 **Format:** Seminar

**Instructor:** Dr. Natalia Umansky
umansky@ipz.uzh.ch
https://nataliaumansky.github.io

**Credits:** 6.0

**Term:** Spring

**Office hours:** available by appointment. **Office:** AFL H349

---

## Module Description

How should democratic societies respond to the amplification of propaganda, disinformation, and hate speech on digital forums designed to promote free expression? Content moderation -the regulation of the material that users create and disseminate online- has become a routine practice as a response to these new challenges presented by the use of social media platforms. However, such practices raise significant questions linked to democratic accountability and civil liberties.

This course will seek to outline the current regulatory practices being employed to control and restrict our online behaviour, and explain the underlying rationales for how, when, and why these policies are enforced. Topics will include the role of algorithms and curation in ranking content; the promise of labeling, fact-checking, and other interventions designed to counter misinformation; and case studies, such as Facebook's Oversight Board.

# Course Programme

| | |
|---|---|
| **20 February** | Introduction and Overview |
| **27 February** | The Promise of Social Media |
| **6 March** | Disinformation and Hate Speech |
| **13 March** | The Myth of the Neutral Platform |
| **20 March** | To Remove, Label, or Filter? Imperfect Solutions at Scale |
| **27 March** | Speech Police: Humans and Machines |
| **3 April** | Platform Governance and Democracy |
| **24 April** | Across-Platform Comparisons |
| **8 May** | What Platforms are and What They Should Be |
| **15 May** | Group Work and Peer-Review: Policy Solutions |
| **22 May** | No class |

# Learning Outcomes

## Knowledge and Understanding

Following this course students will develop a range of important transferable skills.

## Substantive Knowledge

By the end of the course, students should be able to:

- Gain knowledge of the different content moderation practices
- Understand the consequences of platform governance for public discourse
- Understand the current practices employed by different social media platforms
- Explain the underlying rationales for how, when, and why these policies are enforced
- Identify the possible consequences for democracy
- Identify real-life examples of platform governance and its known

## Skills (Intellectual and Transferable)

The course will encourage you to:

- Listen carefully and critically to orally-presented arguments.
- Ability to understand the scientific literature, and in particular to identify research puzzles and knowledge gaps.
- Make links between material presented at different times, on different issues.

- Construct persuasive written, and oral arguments supported by evidence, orally and in writing.
- Read critically and with a clearly defined purpose.
- Apply your theoretical knowledge to the real world.
- Prepare, articulate and defend answers to set questions.
- Formulate and ask your own questions about course material.

The written work in the course will require you to:

- Select relevant material from lectures, literature, news sources, and the web.
- Understand, analyse and assess that material.
- Produce a sustained, structured and informed answer.
- Write in a concise and cogent style.

# Assessment

## Grade Component Breakdown

- Continuous assessment - 30%
  - Individual Presentation - 10%
  - Encounter 5%
  - Group Presentation and Peer-Review 15%
- Policy Memo - 70%

## Individual Presentation - 10%

Over the course of the term, students will need to individually prepare and present one "reaction piece". The short (10 minutes) presentation will engage with at least two of the readings assigned for that week. While you can spend a few sentences summarizing the main points, these presentations should primarily analyze or critique the arguments, identify tensions between them, and suggest constructive ways to synthesize or build on these works.
Students will be asked to sign up to a specific topic/week they would like to cover in their presentation at the beginning of the term. Slides should be submitted on OLAT 3 days **BEFORE** the Monday seminar.

Each student will be required to present at least once in the term. No more than 1 student is allowed to present in a week, and topics will be assigned on a first come first served basis.

## Encounter - 5%

Over the course of the term, students will need to individually prepare and submit one "encounter". In 300/400 words, students will need to describe what they have "encountered" – a song, film, tweet, news story, book, etc. – provide a link (if applicable) and connect it to the material discussed in class. Your work should demonstrate your ability to:

- Identify important, relevant and recent developments.
- Understand and apply the main theoretical approaches covered in the course to analyse real world issues.
- Clearly describe what was "encountered" – a song, film, tweet, video game, book, conversation with a parent, etc. and **connect** it to the theoretical discussions developed in class (ESSENTIAL TO PASS!).
- Be able to explain in few words the relevance of the "encounter" to the topics being discussed in the course.

Students will be asked to sign up to a specific topic they would like to cover in their encounter at the beginning of the term. Encounters should be submitted on OLAT up until the day **BEFORE** the Monday seminar. Encounters CANNOT be submitted for the same topic as the individual presentation.

Each student will be required to submit at least one encounter. No more than 1 student is allowed to submit each week, and topics will be assigned on a first come first served basis.

## Group Presentation and Peer-Review - 15%

In preparation for the final assignment, students will have the opportunity to present their proposed policies to the class. Working in pairs, students will have to provide a thorough (15 minutes) presentation of their policy proposal, explaining why their suggested content moderation regime should be adopted by policy-makers.

Moreover, students will have the opportunity to receive feedback from their peers on how to improve their proposals ahead of the final submission. To this purpose, each group will need to submit on OLAT a short (1500 - 2000 words) draft of their policy proposal BY 8 MAY. The drafts will be circulated with the entire class to allow the audience to prepare comments ahead of the presentations. Moreover, a short (10 minutes) Q&A session will be held after each presentation, allowing the speakers to receive feedback from the audience.

All group presentations will take place on 15 May during the seminar. Besides the draft, groups will have to submit their slides before the seminar taking place on 15 May.

Groups will be formed during the first seminar session and will remain unchanged until the end of the term.

## Policy Memo - 70%

The final written assignment will take the form of a policy memo outlining and justifying a specific policy proposal or content moderation regime to relevant decision-makers. The

memo will be written in pairs and should be 5000 (10% +/-) words long. **Deadline: 1 June**

## Plagiarism

Although this should be obvious, plagiarism – copying someone else's text without acknowledgement or beyond 'fair use' quantities – is not allowed. Plagiarism is an issue we take very serious here in UZH.

Please familiarize yourself with the definition of plagiarism on UZH's website and make sure not to engage in it.

## Late Submission Policy

All written work must be submitted on or before the due dates.

When an extension is necessary, the student will need to contact our Prüfungsdelegierte Naome Czisch (pruefungen@ipz.uzh.ch) BEFORE THE DEADLINE to apply for extenuating circumstances.

## Grades

I am very happy to schedule 1:1 meetings to provide students with further feedback when required. However, students should be advised that grades will not be modified after they are released.

## Participation in class

This course is designed as a seminar. While a short lecture by the instructor will precede the discussion, students are expected to actively participate in class. For that purpose, students will need to follow the assigned readings and come to class ready to engage in dynamic discussions. Moreover, I will sometimes encourage debates by proposing different views and challenging students' arguments. This is not a means of discouraging opposing views or imposing my own perspective on the students. On the contrary, it is a resource I employ in class to invite students to develop critical thinking and learn to construct arguments to support their own perspectives.

## Essay Grading Rubric

The following guidelines should be adhered to when writing your final essay:

- **Statement of Purpose/ Focus and Organisation - 40%**
  - The response is fully sustained and consistently and purposefully focused:
    - Claims are clearly stated, focused, and strongly maintained
    - Claims are introduced and communicated appropriately for the purpose, audience, and task
    - Alternate or opposing claims are clearly addressed
  - The response has a clear and effective organisational structure creating unity and completeness:

    - A variety of transitional strategies is consistently used to effectively clarify the relationships between and among ideas
    - The progression of ideas from beginning to end is logical
    - The introduction and conclusion are effective for audience and purpose
    - Appropriate sentence structure variety produce strong connection between ideas

**Evidence/Elaboration - 40%**

- The response provides thorough and convincing support/evidence for the writer's claim that includes the effective use of sources, facts, and details. The response achieves substantial depth that is specific and relevant:
    - Claims are supported with relevant evidence from credible sources and clear reasoning
    - Use of evidence from sources is smoothly integrated, cited, comprehensive, and concrete
    - A variety of effective argumentative techniques is used
- The response demonstrates strategic use of language to produce clear communication:
    - Precise language clearly and effectively expresses ideas
    - The use of academic and domain-specific vocabulary is clearly appropriate for the audience and purpose

**Editing Conventions - 20%**

- The response displays adequate command of all grade level and preceding level conventions of writing:
    - Some errors in usage and sentence formation may be present, but no systematic pattern of errors is displayed
    - The use of punctuation, capitalisation, and spelling is adequate

# OLAT

Please make sure you have access to the module in OLAT as soon as possible. It is the student's responsibility to make sure that they are signed up to the module correctly and they know how to submit coursework through the appropriate OLAT assignment tab. If you have any issues with OLAT contact the IT Helpdesk to resolve the issue.

Furthermore, module materials such as this syllabus and announcements made outside lectures shall be on OLAT. As such, OLAT is an important communication tool for the module.

# Emails

I will seek to reply to emails within the following 48 hours. However, this might not always be the case. Additionally, I will not reply to emails during the weekend or after working hours.

# Additional Covid-19 Guidelines

Covid-19 continues to pose a threat to our well-being and health. We all need to follow UZH's guidelines. If you are not feeling well, stay home! I will try to make all relevant materials available to everyone using OLAT: I will share the slides after each session and upload all seminar materials.

# Course Reading

## Required Readings:

The following texts shall be used extensively throughout the course, so it is recommended that they are purchased:

- Gorwa, R. (2019). What is platform governance?. Information, communication & society, 22(6), 854-871.
- Flew, T., & Martin, F. R. (2022). Digital Platform Regulation: Global Perspectives on Internet Governance.
- Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media. New Haven: Yale University Press.
- Persily, N., & Tucker, J. A. (Eds.). (2020). Social media and democracy: The state of the field, prospects for reform.
- Jackson, S. J., Bailey, M., & Welles, B. F. (2020). #HashtagActivism: Networks of race and gender justice. MIT Press.
- Welles, B. F., & González-Bailón, S. (Eds.). (2020). The Oxford handbook of networked communication. Oxford University Press, USA.

In addition to these readings, students should keep up to date on current international affairs by reading daily newspapers, or one of the many websites and podcasts devoted to the Global South. This reading is essential as it will allow you to keep up to date with current affairs and identify potential encounter topics. These websites include the following:

- `http://www.foreignaffairs.com`
- `http://blogs.lse.ac.uk`
- CCS Podcast - `https://open.spotify.com/show/0PLCDpeA5KyhPE5JObL5S3?si=Q6QiqSwaTlqczIHc5YqiXw`
- Social Media and Politics podcast - `https://socialmediaandpolitics.org/`

## Detailed Course Programme

### 20 February

### Introduction and Overview

Key readings

- Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. Harv. L. Rev., 131, 1598.
- Keller, D., Leerssen, P. (2020). Facts and where to find them: Empirical research on internet platforms and content moderation. Social media and democracy: The state of the field and prospects for reform, 220, 224.
- Riemer, K., & Peter, S. (2021). Algorithmic audiencing: Why we need to rethink free speech on social media. Journal of Information Technology, 36(4), 409-426.

Further reading

- Alizadeh, M., Gilardi, F., Hoes, E., Klüser, K. J., Kubli, M., Marchal, N. (2022). Content Moderation As a Political Issue: The Twitter Discourse Around Trump's Ban. Journal of Quantitative Description: Digital Media, 2.
- Heldt, A., & Dreyer, S. (2021). Competent third parties and content moderation on platforms: Potentials of independent decision-making bodies from a governance structure perspective. Journal of Information Policy, 11, 266-300.
- Gerrard, Y. (2022). Social Media Moderation: The Best-Kept Secret in Tech. In The Social Media Debate (pp. 77-95). Routledge.

### 27 February

### The Promise of Social Media

Key readings

- Tucker, J. A., Theocharis, Y., Roberts, M. E., & Barberá, P. (2017). From liberation to turmoil: Social media and democracy. Journal of democracy, 28(4), 46-59.
- Etter, M., & Albu, O. B. (2021). Activists in the dark: Social media algorithms and collective action in two social movement organizations. Organization, 28(1), 68-91.

Further reading

- Jost, J. T., Barberá, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., ... & Tucker, J. A. (2018). How social media facilitates political protest: Information, motivation, and social networks. Political psychology, 39, 85-118.
- Gil de Zúñiga, H., Molyneux, L., & Zheng, P. (2014). Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. Journal of Communication, 64(4), 612-634.

- Kim, D. H., Ellison, N. B. (2022). From observation on social media to offline political participation: The social media affordances approach. New Media Society, 24(12), 2614-2634.

- Bimber, B., Cunill, M. C., Copeland, L., Gibson, R. (2015). Digital media and political participation: The moderating role of political interest across acts and over time. Social science computer review, 33(1), 21-42.

- Lindgren, S. (2019). Movement mobilization in the age of hashtag activism: examining the challenge of noise, hate, and disengagement in the #MeToo campaign. Policy & Internet, 11(4), 418-438.

- Lim, M. (2020). Algorithmic enclaves: Affective politics and algorithms in the neoliberal social media landscape. In Affective Politics of Digital Media (pp. 186-203). Routledge.

- Lim, M. (2017). Freedom to hate: social media, algorithmic enclaves, and the rise of tribal nationalism in Indonesia. Critical Asian Studies, 49(3), 411-427.

**6 March**

**Disinformation and Hate Speech**

Key readings

- Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. Social media and democracy: The state of the field, prospects for reform, 10.

- Siegel, A. A. (2020). Online hate speech. Social media and democracy: The state of the field, prospects for reform, 56-88.

Further readings

- Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). Visual mis-and disinformation, social media, and democracy. Journalism & Mass Communication Quarterly, 98(3), 641-664.

- Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020). Combating disinformation in a social media age. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(6), e1385.

- Tandoc Jr, E. C., Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. Journalism, 21(3), 381-398.

- Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. Television New Media, 22(2), 205-224.

- Tucker, J. A. (2023). Computational Social Science for Policy and Quality of Democracy: Public Opinion, Hate Speech, Misinformation, and Foreign Influence Campaigns. Handbook of Computational Social Science for Policy, 381-403.

- Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., ... & Tucker, J. A. (2021). Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. Quarterly Journal of Political Science, 16(1), 71-104.

- Hwang, T. (2020). Dealing with disinformation: evaluating the case for Amendment of Section 230 of the communications decency act. Social Media and Democracy: The State of the Field and Prospects for Reform, 252-285.

**13 March**

**The Myth of the Neutral Platform**

Key readings

- Stewart, E. (2021). Detecting fake news: Two problems for content moderation. Philosophy & Technology, 34(4), 923-940.

- Hallinan, B., Scharlach, R., & Shifman, L. (2022). Beyond neutrality: Conceptualizing platform values. Communication Theory, 32(2), 201-222.

Further reading

- Chen, W., Pacheco, D., Yang, K. C., & Menczer, F. (2021). Neutral bots probe political bias on social media. Nature communications, 12(1), 5580.

- Gerrard, Y. (2020). Social media content moderation: six opportunities for feminist intervention. Feminist Media Studies, 20(5), 748-751.

- DeCook, J. R., Cotter, K., Kanthawala, S., & Foyle, K. (2022). Safe from "harm": The governance of violence by platforms. Policy & Internet, 14(1), 63-78.

- Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. New Media & Society, 22(7), 1266-1286.

- Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. Policy & Internet, 14(1), 79-95.

- Are, C. (2021). The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. Feminist Media Studies, 1-18.

- Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O. (2022). (In) visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. New Media & Society, 14614448221109804.

- Wang, S. (2021). Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. Digital Journalism, 9(1), 64-83.

- Hasinoff, A. A., & Schneider, N. (2022). From Scalability to Subsidiarity in Addressing Online Harm. Social Media+ Society, 8(3), 20563051221126041.

- Wojcieszak, M., Thakur, A., Ferreira Gonçalves, J. F., Casas, A., Menchen-Trevino, E., & Boon, M. (2021). Can AI enhance people's support for online moderation and their openness to dissimilar political views?. Journal of Computer-Mediated Communication, 26(4), 223-243.

**20 March**

**To Remove, Label, or Filter? Imperfect Solutions at Scale**

Key readings

- Yildirim, M. M., Nagler, J., Bonneau, R., & Tucker, J. A. (2021). Short of suspension: How suspension warnings can reduce hate speech on twitter. Perspectives on Politics, 1-13.
- Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press. Ch 7.
- Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. Social Media+ Society, 8(3), 20563051221117552.

Further reading

- Morrow, G., Swire[U+2010]Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. Journal of the Association for Information Science and Technology, 73(10), 1365-1386.
- Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press. Ch 4.
- Keller, D., & Leerssen, P. (2020). Facts and where to find them: Empirical research on internet platforms and content moderation. Social media and democracy: The state of the field and prospects for reform, 220, 224.
- Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11), 4366-4383.
- Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media & Society, 18(3), 410-428.
- Cotter, K. (2021). "Shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. Information, Communication & Society, 1-18.
- Petre, C., Duffy, B. E., & Hund, E. (2019). "Gaming the system": Platform paternalism and the politics of algorithmic visibility. Social Media+ Society, 5(4), 2056305119879995.
- Bode, L., & Vraga, E. (2021). The Swiss cheese model for mitigating online misinformation. Bulletin of the Atomic Scientists, 77(3), 129-133.

**27 March**

**Speech Police: Humans and Machines**

Key readings

- Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2), 2053951720943234.
- Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. ACM Transactions on Computer-Human Interaction (TOCHI), 26(5), 1-35.

Further reading

- Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945.
- Elkin-Koren, N. (2020). Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. Big Data & Society, 7(2), 2053951720932296.
- Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. New Media & Society, 21(7), 1417-1443.
- Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., & Lease, M. (2021, May). The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-14).

**3 April**

**Platform Governance and Democracy**

Key readings

- Gorwa, R. (2019). What is platform governance?. Information, Communication & Society, 22(6), 854-871.
- Gorwa, R., Ash, T. G. (2020). Democratic transparency in the platform society. Social media and democracy: The state of the field and prospects for reform, 286-312.

Further reading

- Duffy, B. E., & Meisner, C. (2022). Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility. Media, Culture & Society, 01634437221111923.
- Banchik, A. V. (2021). Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. New Media & Society, 23(6), 1527-1544.

- Sablosky, J. (2021). Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar. Media, Culture & Society, 43(6), 1017-1042.
- De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. Computer Law & Security Review, 36, 105374.
- MacCarthy, M. (2020). Transparency requirements for digital social media platforms: Recommendations for policy makers and industry. Transatlantic Working Group.

**24 April**

**Across-Platform Comparisons**

Key readings

- Makhortykh, M., Urman, A., Münch, F. V., Heldt, A., Dreyer, S., & Kettemann, M. C. (2022). Not all who are bots are evil: A cross-platform analysis of automated agent governance. New Media & Society, 24(4), 964-981.
- Urman, A., & Makhortykh, M. (2023). How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. Telecommunications Policy, 102477.
- Gillett, R., Stardust, Z., & Burgess, J. (2022). Safety for Whom? Investigating How Platforms Frame and Perform Safety and Harm Interventions. Social Media+ Society, 8(4), 20563051221144315.

Further reading

- Hovyadinov, S. (2019). Toward a More Meaningful Transparency: Examining Twitter, Google, and Facebook's Transparency Reporting and Removal Practices in Russia. Google, and Facebook's Transparency Reporting and Removal Practices in Russia (November 30, 2019).
- Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. Political Communication, 35(1), 50-74.

**8 May**

**What Platforms are and What They Should Be**

Key readings

- Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press. Ch 8.
- Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., ... & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. Internet Policy Review, 9(4), Article-number.

Further reading

- Gehl, R. W., & Zulli, D. (2022). The digital covenant: non-centralized platform governance on the mastodon social network. Information, Communication & Society, 1-17.

**15 May**

Group Work and Peer-Review: Policy Solutions

**22 May**

No class