The Turn to Artificial Intelligence in Governing Communication Online

Gollatz, Kirsten; Beer, Felix; Katzenbach, Christian

Veröffentlichungsversion / Published Version
Sonstiges / other

KIRSTEN GOLLATZ, FELIX BEER, CHRISTIAN KATZENBACH

The Turn to Artificial Intelligence in Governing Communication Online

Workshop Report

HIIG WORKSHOP REPORT 2018-09

ABSTRACT

Presently, we are witnessing an intense debate about technological advancements in artificial intelligence (AI) research and its deployment in various societal domains and contexts. In this context, media and communications is one of the most prominent and contested fields. Bots, voice assistants, automated (fake) news generation, content moderation and filtering – all of these are examples of how AI and machine learning are transforming the dynamics and order of digital communication.

On 20 March 2018 the Alexander von Humboldt Institute for Internet and Society together with the non-governmental organisation Access Now hosted the one-day expert workshop "The turn to AI in governing communication online". International experts from academia, politics, civil society and business gathered in Berlin to discuss the complex socio-technical questions and issues concerning subjects such as artificial intelligence technologies, machine learning systems, the extent of their deployment in content moderation and the range of approaches to understanding the status and future impact of AI systems for governing social communication on the internet.

This workshop report summarises and documents the authors' main takeaways from the discussions. The discussions, comments and questions raised and responses from experts also fed into the report. The report has been distributed among workshop participants. It is intended to contribute current perspectives to the discourse on AI and the governance of communication.

KEYWORDS

Content moderation, machine learning, platform governance, society-in-the-loop, artificial intelligence, social media

DISCLOSURES

All authors work or have previously worked with the Alexander von Humboldt Institute for Internet and Society in Berlin.

The expert workshop and the publication of this report was supported by a grant from the Volkswagen Foundation, Hanover, Germany.

CONTENTS

1 INTRODUCTION **3**

2 THE PROGRAMME OF THE DAY **4**

3 REPORT ON THEMATIC SESSIONS **6**

Session 1: Detecting and Classifying Content 6

Session 2: Humans and Machines – Division of Labour and Practices 9

Session 3: Policy and Governance Instruments 12

Session 4: AI and Society-in-the-Loop: Societal Implications 15

4 THE PATH AHEAD **17**

5 REFERENCES **18**

ANNEX: LIST OF PARTICIPANTS **19**

1 INTRODUCTION

"Artificial Intelligence" and "Machine Learning" systems are transforming and reorganising various spheres of society. Content moderation and communication governance on digital platforms have emerged as a prominent, but increasingly contested field of application for automated decision-making systems.

Major technology companies such as Facebook, Twitter, YouTube, or Google are shaping the communication ecosystem in large parts of the world. Smartphones and tablets, search engines and social media have by and large replaced traditional media as primary gateways to information. While this has created new opportunities for people to connect in various ways around the world, it also offers opportunities for users to upload content that is objectionable, including images of child abuse and gratuitous violence, as well as disturbing, hateful messages.

Particularly with the rapidly growing political impact of misinformation and hate speech, there have been increasing calls for online platforms to prevent and remove problematic content. Governments around the world have initiated regulatory policies to restrict online speech they deem to be unlawful, among them France, Vietnam, Russia, Singapore and Venezuela. The German "Netzwerkdurchsetzungsgesetz" (Network Enforcement Act), for example, constitutes one such attempt to improve national law enforcement on platforms by requiring their operators to swiftly detect and take down content that is defined as unlawful under German law. In the absence of commonly agreed upon speech norms and coherent regulatory frameworks, these regulatory policy initiatives by governments must be seen in the context of transnational challenges of cross-border content regulation (Gollatz, Riedl & Pohlmann, 2018).

In the context of this policy pressure, and with the massive amount of content uploaded to platforms every day, companies have turned towards greater automation of content moderation. They routinely present machine learning technologies as catch-all solution to detect and filter hate speech, misinformation, and copyright infringements. However, this expansion of algorithmic decision-making brings its own set of problems. The opaque implementation, vague definitions and lack of accountability of these AI systems can cause problems such as overblocking or biased decision making. Experts and activists warn that a hasty implementation of AI driven solutions may have detrimental effects on the freedom of speech and equal access to information on the internet.

2 THE PROGRAMME OF THE DAY

On March 20, 2018, the Alexander von Humboldt Institute for Internet and Society and Access Now organised the one-day transdisciplinary workshop "The turn to artificial intelligence in governing communication online". International experts from academia, politics, civil society and business gathered in Berlin to discuss technological advancements, the extent of artificial intelligence deployment and the range of approaches to understanding the status and future impact of AI-systems for governing communication on the internet.

Workshop Programme of March 20, 2018

Time	Session	Thematic Scope
9:00 am	Welcome and Introductions	
9:45 am	Kick-off Statements	- Malavika Jayaram: *Napalm, Nuance and Not Hot Dogs* - on practices of balancing the fight against hateful content and protection free speech in Asia - Nick Feamster: *AI and the Future of Free Expression Online* - on manipulation and filtering of platforms
10:30 am	Session 1: Detecting and Classifying Content	- Detection of objectionable content and its technical implementation - Drivers of change and limitations of AI
12:00 pm	Session 2: Humans and Machines – Division of Labour, Practices	- Content moderation practices involving humans AND machines - Mitigation of human error but also evidence for the need for more human decision-making
2:00 pm	Session 3: Policy and Governance Instruments	- General framing of AI and governance of platforms - Debates and regulatory initiatives on EU level - Concrete applications and regulatory interventions
3:30 pm	Remote Intervention by Tarleton Gillespie	- Content moderation as an essential commodity of platforms - Use of AI for democratising platform governance
3:45 pm	Session 4: AI and Society-in-the loop	- Bringing society into the discussion, its perceptions and expectations towards AI - Ethics in society-aware design of ML-systems
5:00 pm	Remote Intervention by David Kaye	- To what extent are automated rules framing what people are able to speak and receive, or think about global free speech norms?
5:30 pm	Wrap Up and the way forward	- The paradox of power without responsibility - Appropriate/inappropriate definitions of content - Translations into technically designs

* Some experts participated remotely.

In four sessions, the participants explored different themes surrounding the implications of AI in moderating content online.

The diversity of participants' backgrounds, knowledge and expertise has been a major advantage in this workshop. It reflected the complexity and variety of perspectives that are pertinent in the emerging public and expert debates around AI and the governance of online communication. The workshop particularly aimed at transcending disciplinary boundaries and also involved non-academics from business, policy and civil society.[1]

By bringing together a range of different stakeholders, the workshop served as an interdisciplinary vantage point for the participants to share their expertise and discuss their perspectives on the subject matter. In the following, the relevant insights and takeaways of all workshop sessions are thematically presented.

[1] Please refer to the Annex for a list of participating experts.

3 REPORT ON THEMATIC SESSIONS

Session 1: Detecting and Classifying Content

Scope: The first session centred around the question of how we can observe and describe the current turn towards AI-driven solutions for detecting and classifying problematic content on online platforms. The discussion dissected the rationale behind this development and explored the technologies' capabilities and limitations as well as larger societal implications.

With contributions by: Renata Barreto (UC Berkeley), Sabine Frank (Google), Emma Llansó (Center for Democracy & Technology), Fabrizio Augusto Poltronieri (De Montfort University), Betty van Aken (Beuth University) and Zeerak Waseem (University of Sheffield)

Moderator: Christian Katzenbach (HIIG)

The starting point of the workshop was a discussion about the driving factors behind online platforms' increased attention to AI systems in content moderation. Participants agreed that a key factor in this development certainly is the mounting public pressure to take swift action against problematic content in recent times. The perceived prevalence of misinformation and hate speech in recent political events such as Brexit, the election of Donald Trump and more generally the surge of right wing populism has created a widespread sense of urgency about inadequate law enforcement and regulation in the digital public sphere. New legislative initiatives are pressuring online platforms to depart from their formerly rather neutral stance and engage proactively with problematic content.

COMPARED TO HUMAN CONTENT REVIEWERS, THE ADVANTAGE AND APPEAL OF AUTOMATED SYSTEMS IS THEIR SCALABILITY

This means that platforms have to detect, classify and evaluate an enormous quantity of uploaded content day by day. Compared to human content reviewers, the advantage and appeal of automated systems is clearly their scalability. Such systems promise to make the content moderation process much easier, quicker and cheaper than would be the case when using human labour.

While automated filtering is currently mainly used to complement the work of human content reviewers, the industry has high expectations regarding the ability of automated decision making to replicate the nuanced judgement of human moderators in the foreseeable future. Despite big corporate and public investments into the research and development of AI applications, some experts consider these expectations unrealistic and overly optimistic. They argue that the push for automated content moderation systems has to be seen within a larger atmosphere of AI enthusiasm. Contrary to this technological solutionism, workshop participants highlighted the dangers and downsides that automated online communication filtering may bring about. Some concerns that were raised in this context included overbroad censorship,

infringements on speech and associations rights, and biased decision making against minorities and non-English speakers.

Afterwards, the session's focus shifted towards examining the capabilities and limitations of current content moderation practices (Duarte, Llansó & Loop, 2018). Automated content filtering is not new. Over the years, many tools have been deployed to analyse and filter content, including tools for spam detection or hash matching. These tools identify unwanted content on the basis of certain sharply defined criteria derived from previously observed keywords, patterns or metadata.

The effectiveness of automated social media moderation tools, however, largely depend on their ability to accurately analyse and classify content in its context. The capability to parse the meaning of a text is highly relevant for making important distinctions in ambiguous cases, i.e. when differentiating between hate speech and irony. For this task, the industry has now increasingly turned to machine learning to train their programmes to become more context sensitive.

The participants then discussed the capabilities and limitations of this current approach. The success of automated content moderation systems is usually evaluated in terms of accuracy rates, which give an indication of how closely their judgement matches human decision-making on average. To achieve high accuracy rates, algorithmic training must focus on a clearly defined type of data. This means that automated classification and detection systems are usually specifically trained to evaluate these cases and are therefore not transferable to other domains. However, the more ambiguous and contextual classificatory criteria become, the more difficult it becomes to train algorithms accurately. On the one hand, AI solutions are an effective tool to filter clearly confined cases such as child pornography. On the other hand, even humans struggle with making consistent judgements in certain cases – for example, when drawing clear distinctions between political activism and calls for violence – and automated systems are far behind humans in this respect.

AUTOMATION WILL LIKELY LEAD TO A SHIFT FROM REACTIVE TO PROACTIVE MODERATION WHICH WOULD MAKE ACCOUNTABILITY, TRANSPARENCY AND PUBLIC PARTICIPATION VITALLY IMPORTANT

AI systems lack (at least for now) human beings' language sensibility and understanding of semantics, which is required for this difficult task. Several participants explained that most of today's filtering techniques boil down to flagging content based on certain keywords. The meaning of language is highly context sensitive and constantly in flux; a word could radically change its meaning if used at different places over time. A content moderation system that bases its classifications simply on certain keywords cannot attain this level of complexity and runs the risk of producing unexpected false positives and negatives in the absence of context.

To avoid structural overblocking, human involvement consequently remains an essential part of content moderation, at least in highly context sensitive cases to avoid structural overblocking. Beyond this, some concerns were raised about whether automation could lead to a further marginalisation of groups that already face discrimination because of social biases and errors inherent in the training data. Online platforms must also take these implications into account in the design process of their AI systems.

Building on this, the discussion questioned under which circumstances automation can be deemed useful. This question not only depends on the type of content to be verified. The deployment of automated systems varies also with regard to the different stages in the process of content moderation. Pre-moderation, post-moderation, reactive moderation or distributed moderation were just some of the concepts discussed in this context. All forms of automation carry their individual chances and risks. There was large agreement that automation would most likely lead to a shift from reactive (i.e. triggered by user flagging) to proactive moderation (i.e. analysing all uploaded content). Participants articulated that this scenario raises serious concerns and exacerbates existing problems of accountability and lack of transparency. In response to this, the workshop discussed options to democratise the overall design and deployment of content moderation by ensuring the intelligibility of the process and allowing for proper public participation.

Session 2: Humans and Machines – Division of Labour and Practices

Scope: The second session explored what the division of labour between humans and machines will look like in the future of content moderation. In this session, the participants discussed the possibilities for AI to replace or assist human labour in the content moderation process.

With contributions by: Johannes Baldauf (Consultant), Ulrike Klinger (Freie Universität Berlin, Weizenbaum Institute), Iva Nenadic (European University Institute), Sarah T. Roberts (UCLA), Jeremy Rollison (Microsoft), Mirko Vossen (die medienanstalten) and Jillian C. York (EFF)

Moderator: Kirsten Gollatz (HIIG)

Claims that AI systems will make humans obsolete are commonplace. This black and white reasoning is challenged by many experts, who explain that automation will transform rather than replace human labour. The same holds true for online content moderation. There's no question that the way online platforms monitor content and remove offensive material is on the brink of change. Contrary to the belief that AI will entirely supplant human review, there was strong consensus among our attendees that effective moderation will have to rely on a hybrid model in the foreseeable future: while some tasks are highly amenable for automation such as identifying sentences that clearly infringe community guidelines or pre-selecting suspicious cases in large quantities of data in near real time, others will continue to require human judgement – i.e. the use of contextual knowledge to decide on grey areas. In short, humans and machines will rely on a synergistic relationship. It was widely held that AI should be merely an assistive technology to increase scalability and improve human effectiveness and efficiency when judging content.

AI SYSTEMS IN CONTENT MODERATION WILL NOT MAKE HUMANS OBSOLETE BUT WILL RATHER TRANSFORM THE JOB OF A CONTENT MODERATOR

Figuring out the appropriate division between machines and humans will thus be a challenging task. At the moment, most major online platforms contend that they only employ automation for flagging content, and that the final removal decision is left to human reviewers.[2] However, content moderation is still dependent on cohorts of low-skilled labourers, mostly employed by subcontractors in India and the Philippines, with wages well below the average Silicon Valley tech employee. These moderators spend their days reviewing vast amounts of content to decide on whether it should be removed or not, applying appropriateness criteria that are often ambiguous and opaque (Arsh & Etcovitch, 2018).

Our discussion drew attention to the growing body of evidence suggesting that content moderation in its

[2] See, for example: YouTube Official Blog. (2017, 4 December). Expanding our work against abuse of our platform. Retrieved from https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html

current form exposes employees to considerable psychological risks (Gillespie, 2018).[3] Many of these individuals are required to meet intolerable numerical quotas, as they screen disturbing content such as beheading, suicide or pornographic videos. Investigative journalists stress that PTSD-like symptoms and other mental health issues often arise as a consequence among moderators. The attendees highlighted the fact that an accurate evaluation of the existing systems is being hampered as online platforms provide only very little information about the issue. The companies are often intentionally opaque, resist any attempt by third parties to monitor their practices and use non-disclosure agreements prevent employees from discussing their working conditions.[4]

Many observers have problematised this model, but few have proposed alternative solutions. AI might, however, open a window of opportunity for a new, more synergistic labour division between humans and machines. Semi-automated moderation models have the potential to replace large numbers of low-skilled positions, with a new role for human reviewers augmented by AI. The next generation of AI tools will be able to identify and evaluate content on a larger range of attributes than what is currently possible, including content source and context. Based on this, a relative risk score can be calculated to determine whether something should be posted immediately, after review, or not at all.

With machine learning, results can be used to optimise algorithms autonomously to continuously enhance their accuracy. This process might eliminate large volumes of content from the investigator's queue and could allow content moderators to concentrate on decision making in complex grey areas. They can bring specialised expertise, empathy, and contextual knowledge to judge these specific types of content. This would mean that we will most likely see today's content moderators being supplanted by content investigators with specialist training akin to that of a financial crime investigator.[5] These content moderators could be trained in language, regional, market, regulatory and legal specificities to make well-informed decisions when it comes to grey area content. This transformation has the potential to help rapidly growing online platforms to scale content moderation at affordable costs while minimising risk and significantly improving the career prospects of human reviewers.

AUTOMATED MODERATION TOOLS NEED TO BE FUTHER IMPROVED IN THE FUTURE SO THAT THEY LOWER MODERATORS' EXPOSURE TO DISTURBING CONTENT

Despite this, many hybrid models still face considerable challenges and unresolved issues. While increasing the efficiency and scalability of content moderation seems attainable, pre-existing problems such as intransparency or lacking accountability remain largely unresolved and may even be further aggravated

[3] See also: Chen, A. (January 28, 2017), The Human Toll of Protecting the Internet from the worst of Humanity, The New Yorker. Retrieved from https://www.newyorker.com/tech/elements/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity; Buni, C. & Chemaly, S. (April 13, 2016), The Secret Rules of the Internet, The Verge. Retrieved from https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech

[4] Powers, B. (Sept. 9, 2017), The Human Cost of Monitoring the Internet, The Rolling Stone. Retrieved from https://www.rollingstone.com/culture/features/the-human-cost-of-monitoring-the-internet-w496279

[5] See also: Accenture: Content Moderation: The Future is Bionic. Retrieved from https://www.accenture.com/cz-en/_acnmedia/PDF-47/Accenture-Webscale-New-Content-Moderation-POV.pdf

through partial automation.[6] For example, there were various concerns that automation could further increase the opaqueness of the moderation process and proactively contribute to over-policing content. Although improved automated pre-selection tools could lower overall exposure to disturbing content, moderators will probably still have to deal with vast amounts of such material. Hence, companies must step up their efforts to create a healthy working environment for their moderators.

[6] See also: Guiding Principles for the Future of Content Moderation: Four Scholars and Advocates in Conversation, February 1, 2018. Retrieved from https://atm-ucla2017.net/

Session 3: Policy and Governance Instruments

Scope: The third session dealt with policy and governance instruments for regulating communication online in the context of the turn to AI. At first, the participants debated the regulatory status quo and the rationale and implications of the current push for increased regulation. Looking at what has and has not worked so far, the discussion then examined the possibilities and challenges for improved regulation models.

With contributions by: Prabhat Agarwal (EU Commission), Eimear Farrell (Amnesty International), Amélie Heldt (Bredow Institute), Michael Latzer (University of Zürich), Ramak Molavi (iRights), Erin Saltman (Facebook, remotely), Florent Thouvenin (University of Zurich) and Joris van Hoboken (Vrije Universiteit Brussels)

Moderator: Fanny Hidvégi (Access Now)

Until now, digital platforms have benefited from little scrutiny and reduced governmental interference, leaving them vast leeway with regard to the implementation of content moderation measures. This long-standing status quo rests on the binary premise that platforms either function as news publishers that carry full responsibility for their content, or as neutral intermediaries without any legal liability (Tushnet, 2008). In general, the latter model has been seen as the preferable approach because it was believed to reinforce sectoral innovation and freedom of expression in the digital sphere.

With mounting criticism over problematic content, online platforms and policy makers alike face increasing pressure to act, which has led to a new cross-sectoral consensus: platforms can and should do more to regulate their content. Coinciding with increasing attention to AI solutions, this pressure has guided the current algorithmic turn in content moderation processes.

THERE IS PRESSURE TO ACT BUT ALSO A NEW CROSS-SECTORAL CONSENSUS: PLATFORMS CAN AND SHOULD DO MORE TO REGULATE THEIR CONTENT, ALSO WITH THE USE OF AI SYSTEMS

Content moderation has to strike a delicate balance between creating a safe online environment and ensuring free speech. Some believe that governments should take action to regulate their services and energetically fight the uncontrolled spread of problematic content. Others argue that platforms should better be left to regulate content themselves to avoid adverse effects on business innovation and freedom of speech. What form of governance model is therefore probable and preferable for regulating online communication in the future?

In recent years, social media platforms have increasingly committed to voluntary codes of conduct to fight problematic content and have stepped up their efforts to proactively prevent and remove such content. For

instance, in December 2017, Twitter implemented new policies to prevent harassment and hate speech,[7] YouTube is adding additional human content moderators and expanding flagging algorithms,[8] and Facebook also plans to increase their content moderator staff up to 20,000 this year and CEO Mark Zuckerberg declared fixing abuse on Facebook even his personal goal for 2018.[9] While this progress is widely lauded, critics suspect that these efforts will not suffice to create a healthy and safe online communication ecosystem.

In addition, governments increasingly mandate and incentivise more rigorous content moderation. The problem with hard legislation is that it could easily lead to over-policing of content and thus drastically curtail the freedom of speech in the digital sphere. For example, under the German NetzDG, social networks could be punished with up to 50 million euros (approx. USD 60 million) in fines if identified illegal hate speech is not removed within 24 hours. This may lead online platforms to overreact and to speed up the implementation of opaque and imprecise automated content moderation systems to avoid punishment. Despite governmental involvement in platform-content relations, these approaches still rely heavily on platforms' self-regulatory mechanisms to develop the AI systems necessary for the identification and removal of problematic content.

Workshop participants concluded that adequate governance models will rely on a balanced mix of public and private interventions. Generally, participants agreed that government intervention is only desirable to the extent that self-regulation is incapable of dealing with the moderation of problematic content. Hence, there should be as much government regulation as needed but as little as possible. Of course, there is much debate about what this in fact means. How can we regulate content flows without the hampering growth and innovation of online platforms? How can governance models reflect public concerns regarding the transparency, legitimacy and accountability of automated content moderation? In close dialogue with other stakeholders, policy makers should continuously ask themselves questions like these to assess the strengths and shortcomings of state regulation and self-regulatory regimes and by this to determine their role in the process of automated content moderation.

CONTEXT-SPECIFIC DIFFERENCES OF SPEECH ACTS IN DIFFERENT COUNTRIES AND CULTURES BECOME EXTREMELY COMPLICATED WHEN APPLIED TO AUTOMATED DETECTION OF PROBLEMATIC CONTENT

Moreover, there was also a discussion about the complex interaction between different actors in the governance of online communication. Parallel to many other domains, in the digital sphere formal authority has been dispersed from central states to supranational institutions, multinational tech giants and the globalised digital civil society. Reconciling these various stakeholder interests creates a high degree of complexity. Therefore, if the goal is to achieve an effective and global regulatory regime for online communication, at which territorial and institutional levels is decision-making possible?

[7] https://blog.twitter.com/official/en_us/topics/company/2017/safetypoliciesdec2017.html

[8] https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html

[9] https://www.facebook.com/zuck/posts/10104380170714571?pnref=story

This is further complicated by the fact that different countries have varying laws and cultural restrictions on speech. For example, "incitement to violence" or "hate speech" refer to a different thing in Germany – where Nazi propaganda is illegal – than in Spain – where it is illegal to insult the king. These context-specific differences become extremely complicated when applied to automated detection of problematic content. Already these quite simple differences show that training a single content classifier for universal use is impossible. In this case, each jurisdiction would essentially require a different content classifier. A common framework for content moderation can therefore not be fixed but must be continuously negotiated with all involved stakeholders. This confronts policy makers with a significant challenge: an international governance model must be specific enough to exercise regulatory power, but also adaptive enough to allow for context-dependant nuances.

A COMMON FRAMEWORK FOR CONTENT MODERATION CANNOT BE FIXED BUT MUST BE CONTINUOUSLY NEGOTIATED WITH ALL INVOLVED STAKEHOLDERS

In conclusion, our participants agreed that online communication governance should be a multi-stakeholder and multi-level process aimed at developing global guidelines and exchanging best practices to address the challenges of automated content moderation. As this process is still in an early stage of its development, our participants pointed out a number of obstacles that hamper the advancement of effective governance:[10] First, there is a lack of institutional procedures and platforms to facilitate the cross-sectoral conversation about this topic. Second, online platforms currently control all aspects of content moderation and are very secretive about their governance models. Particularly their monopoly on relevant data and information makes it difficult to audit the instruments that platforms are currently employing. Thirdly, without access to reliable data, it is very difficult to engage in joint research on this topic. Lastly, with very little knowledge and resource sharing in place, it is particularly difficult for smaller businesses to develop and train their own algorithmic moderation systems from scratch.

[10] For further reading see: NYU Stern Center for Human Rights and Business (Nov 2017). Harmful Content: The Role of Internet Platform Companies in Fighting Terrorist Incitement and Politically Motivated Disinformation, White Paper. Retrieved from http://www.stern.nyu.edu/experience-stern/faculty-research/harmful-content-role-internet-platform-companies-fighting-terrorist-incitement-and-politically

Session 4: AI and Society-in-the-Loop: Societal Implications

Scope: The fourth session shifted the discussion to societal implications of automated content moderation processes. While the previous discussion focused on the capabilities and limitations of different governance models, this session raised the question of how algorithmic authority can be designed in a transparent, democratic, and accountable way.

With contributions by: Amar Ashar (Berkman Klein Center), Lisa Gutermuth (Ranking Digital Rights), Aphra Kerr (Maynooth University), Tilo Mentler (University of Lübeck), Kevin Morin (Institut National de la Recherche Scientifique), Jörg Pohle (HIIG) & Matthias Spielkamp (Algorithm Watch)

Moderator: Christian Katzenbach (HIIG)

On the one hand, online platforms face mounting pressure to increase in-house regulation of content. On the other hand, human rights defenders and activists are voicing concerns that automated processes will facilitate over-policing of content and lead to erroneous decision making. Automated censorship in the form of takedowns, blocking and filtering content is increasingly implemented by online platforms as a response to problematic expressions. This approach threatens to infringe on individual's freedom of expression and to disproportionately impact groups who already face discrimination in society; in other words, groups who utilise social media to amplify their voices, form associations, and organise for change.

THERE IS AN URGENT NEED TO NOT ONLY DEBATE THE ADVERSE IMPLICATIONS OF AUTOMATED CONTENT MODERATION PROCESSES, BUT ALSO TO EXPLORE AI DESIGNS THAT PUT SOCIETY IN THE LOOP

To holistically address diverse societal interests, these algorithmic systems need to adhere to democratic standards of transparency and accountability, contestation and participation, and include appeals and remedies. For this reason, there is an urgent need to not only debate the adverse implications of automated content moderation processes, but also to explore AI designs that put society in the loop (Rahwan, 2018; Link et al., 2016).

Defining "problematic content" or not?

The conceptual idea of society-in-the-loop raised the question about the role that society should and can play in online content moderation processes. Our participants started by discussing whether or not online platforms should unilaterally define what content is acceptable. Until now, companies set the reference framework for their services independently through their terms of service regulations.

Need for transparency, but on what, and for whom?

Right now, online platforms operate with little scrutiny and often even decide to make their practices

opaque for external observers (O'Neil, 2016). Transparency vis-à-vis governments and independent researchers is however essential to help society to understand the consequences of platform's content moderation practices. Citizens need to know how these platforms operate, how they shape user experiences, and what the companies are doing with their content. For this reason, many experts demand access to their data and systems. Although this matter is highly complicated for various reasons, it is necessary to ensure that society profits from these technologies in the long-term. Our participants recommended to open up new conversation channels centered around how to create frameworks and mechanisms that provide for such multistakeholder scrutiny.

WE NEED TO OPEN UP NEW CONVERSATION CHANNELS CENTERED AROUND HOW TO CREATE FRAMEWORKS AND MECHANISMS THAT PROVIDE FOR MULTISTAKEHOLDER SCRUTINY

Human rights impact assessment, and assigning responsibilities in automated decision-making

Given that companies are constantly introducing new products, updating their policies, and expanding into new jurisdictions, human rights impact assessments should be carried out on an ongoing basis, and should not be a one-time event (UN Human Rights Council, 2018). Human rights impact assessments should include all human rights that companies' policies may impact, beyond freedom of expression and privacy, to include also economic, social and cultural rights, the right to be free from violence, and the right to participate in public life, among others. In addition, they should consider how their policies can strengthen, rather than undermine, due process.

Remedy mechanisms and due process

Governance models should be able to not only evaluate the performance of online platforms, but also allow for the adoption of tailored remedies.

The realisation of this model may raise certain practical challenges and problems. However, we believe that these should be further debated and explored. In today's digital societies, we must engage collectively around the emerging issues that fundamentally concern the right to freedom of expression and human rights more broadly.

4 THE PATH AHEAD

The discussion at the Berlin workshop has mapped and explored the problem space, its levels of complexities and angles. We have distributed the report among participants and published it on our website. In addition to our project website we regularly tweet about issues and interact with a growing community under the hashtag #Turn2AI[11] on Twitter.

FOLLOW AND ENGAGE IN THE DISCUSSION! Use #Turn2AI

For the future, we hope to maintain a network of academic and non-academic experts in this field. The workshop served as a forum for a fruitful exchange both between academic disciplines as well as between academics, civil society and practitioners with diverse backgrounds and interests. For example, members of the workshop have already hosted a panel discussion at RightsCon in Toronto[12] on May 17, 2018 titled "This Panel May Contain Sensitive Content: Automated Filtering and the Future of Free Expression Online".

We will also publish a special issue at the high-ranking journal *Big Data & Society* based on the expert workshop in terms of topic and content. The special issue will include interdisciplinary academic articles that address the role of AI in governing communication online. This includes – but is not limited to – contributions that direct attention to the following questions:

- What factors and actors are driving this change towards more automation and AI in content moderation and regulation on social media platforms?
- What are the technical opportunities and limitations in developing and use of AI in communication governance?
- What are the social and legal expectations towards this technology? Do these expectations have an impact on software development? If so, how?
- How can future AI systems be developed and trained? Is it actually possible to optimise them for the public good?

As we embark on this critical and forward-looking field of work, we hope to maintain this emerging network of experts and plan to contribute to the discourse on AI and communication governance within our research and advocacy networks in the future.

[11] https://twitter.com/hashtag/Turn2AI

[12] RightsCon is a summit series that convenes the global community on human rights and technology. It is hosted by Access Now. RightsCon Toronto took place from May 16 to May 18, 2018 in Toronto, Canada.

5 REFERENCES

Arsht, A. & Etcovitch, D. (March 2, 2018). The Human Cost of Online Content Moderation, Harvard Law Review Online, Harvard University, Cambridge, MA, USA. Retrieved from https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation

Duarte, N., Llansó, E. & Loup, A. (2018): Mixed Messages? The Limits of Automated Social Media Content Analysis, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:106-106. Retrieved from http://proceedings.mlr.press/v81/duarte18a.html

Gillespie, T. (2018). *Custodians of the Internet : Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, London: Yale University Press.

Gollatz, K., Riedl, M. J. & Pohlmann, J. (August 9, 2018). Removals of Online Hate Speech in Numbers. HIIG Science Blog, Alexander von Humboldt Institute for Internet and Society, Berlin, Germany. Cross-posted at Media Policy Project Blog, London School of Economics, London, UK. Retrieved from https://www.hiig.de/en/removals-of-online-hate-speech-numbers/ and http://blogs.lse.ac.uk/mediapolicyproject/2018/08/16/removals-of-online-hate-speech-in-numbers/

UN Human Rights Council (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, on the regulation of user-generated online content. Human Rights Council Thirty-eighth session from 18 June–6 July 2018. Retrieved from http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35

Link, D., Hellingrath, B. & Ling, J. (2016). A Human-is-the-Loop Approach for Semi-Automated Content Moderation, Long Paper – Social Media Studies Proceedings of the ISCRAM 2016 Conference – Rio de Janeiro, Brazil, May 2016. Retrieved from https://pdfs.semanticscholar.org/2223/f7245f7c310db2c4a24e6e4603c85936d460.pdf

O'Neil, C. (2016). *Weapons of Math Destruction. How big data increases inequality and threatens democracy*. New York: Crown Books.

Rahwan, I. (2018). Society-in-the-Loop. Programming the Algorithmic Social Contract, Rahwan, I. Ethics Inf Technol (2018) 20: 5. https://doi.org/10.1007/s10676-017-9430-8

Tushnet, R. (2008). Power Without Responsibility: Intermediaries and the First Amendment, 76 Geo. Wash. L. Rev. 101. Retrieved from http://scholarship.law.georgetown.edu/fwps_papers/76

ANNEX: LIST OF PARTICIPANTS

(in alphabetical order)

Prabhat Agarwal European Commission, Belgium

Amar Ashar Berkman Klein Center for Internet & Society at Harvard University, United States

Johannes Baldauf Consultant, Germany

Renata Barreto University of California Berkeley, United States

Eimear Farrell Amnesty International, Germany

Nick Feamster Princeton University, United States

Sabine Frank Google, Germany

Tarleton Gillespie Microsoft Research New England, United States (remotely)

Kirsten Gollatz Alexander von Humboldt Institute for Internet & Society, Germany

Lisa Gutermuth Ranking Digital Rights at New America, United States

Amélie Heldt Hans-Bredow-Institut ; Alexander von Humboldt Institute for Internet & Society, Germany

Fanny Hidvégi Access Now, Belgium

Jeanette Hofmann Berlin Social Science Center, Germany

Malavika Jayaram Digital Asia Hub, Hong Kong

Christian Katzenbach Alexander von Humboldt Institute for Internet & Society, Germany

David Kaye UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (remotely)

Aphra Kerr Maynooth University, Ireland

Ulrike Klinger Freie Universität Berlin ; Weizenbaum Institute for the Networked Society, Germany

Michael Latzer University of Zurich, Switzerland

Emma Llansó Center for Democracy & Technology, United States

Tilo Mentler University of Lübeck, Germany

Ramak Molavi iRights.Law, Germany

Kevin Morin Institut National de la Recherche Scientifique, Canada

Iva Nenadic European University Institute, Italy

Jörg Pohle Alexander von Humboldt Institute for Internet & Society, Germany

Fabrizio Augusto Poltronieri De Montfort University, United Kingdom

Sarah T. Roberts University of California Los Angeles, United States (remotely)

Jeremy Rollison Microsoft, Belgium

Erin Saltman Facebook, United Kingdom (remotely)

Björn Scheuermann Humboldt-Universität zu Berlin ; Alexander von Humboldt Institute for Internet & Society, Germany

Matthias Spielkamp Algorithm Watch, Germany

Florent Thouvenin University of Zurich, Switzerland

Betty van Aken Beuth University of Applied Sciences, Germany

Joris van Hoboken Vrije Universiteit Brussels, Belgium

Mirko Vossen die medienanstalten, Germany

Zeerak Waseem University of Sheffield, United Kingdom

Jillian C. York Electronic Frontier Foundation ; Center for Internet and Human Rights, Germany