**Organization for Security and Co-operation in Europe**
**Office of the OSCE Representative on Freedom of the Media**

# SAIFE Expedition

15th – 16th December 2022

# ANNOTATED AGENDA

## ANNOTATED AGENDA

## Friday, 16th December 09.00 - 17.30 CET

| Room 531 | | | |
|---|---|---|---|
| **Time** | **Title** | **Speakers** | **Description** |
| **09.00-10.00** | **Datafication of our lives** | **Julia Haas** (OSCE RFoM), **Peter Kirchschläger** (University of Luzern), **Cosima Terrasse** (Laokoon) | In this interdisciplinary session of Art and Ethics, concrete case studies of the datafication of our lives threatening freedom of opinion will be analyzed. This exchange will be enriched by elements of the interactive storytelling documentary "Made to Measure". "Made to Measure" is an experiment that shows that you can reconstruct a person's biggest vulnerabilities solely with their digital data trail – and how social engineering can lead to significant manipulation. The session will discuss what "datafication" means, the process from data to user profiles, and how this shapes our political and public discourse. Finally, possible solutions how to address these challenges will be discussed, among others, human rights-based data-based systems as well as regulatory responses for the design, development and deployment of digital technologies and data-based systems. |
| **10.00-10.30** | **Anatomy of an AI system** | **Vladan Joler** (SHARE Foundation) | This session will explore three large-scale artistic maps of AI systems (Nooscope, New Extractivism, and Anatomy of an AI System). The view diagrams combine and visualize central, extractive processes that are required to run a large-scale artificial intelligence system: material resources, human labor, and data. During the session, the artist Vladan Joler will explore with the audience the vast networks that underpin the "birth, life, and death" of advanced machine-learning technologies, and analyse the limits of AI, and the impact its use has on human perception, autonomy and rights, in particular freedom of opinion, expression and information. |
| **10.30-11.00** | **Society by Design: AI and freedom of expression in local context** | **Sanja Stankovic** (OSCE Mission to Serbia) | The impact of artificial intelligence on freedom of expression is a global topic that requires a broad platform for discussion and action in local environments. With digitalization among the key government's priorities, Serbia has recognized the value of machine learning and AI for the overall development. Within its broader support to democratic reforms in the country, the OSCE Mission to Serbia has been bringing new topics relevant for freedom of expression and media - including AI - closer to citizens and authorities alike.<br>The Mission, together with the RFoM, also focused on bringing together relevant actors from all sectors to address main issues in the intersection of advanced technologies and human rights, in particular freedom of expression. This session will present key takeaways from the joint Conference *Society by Design – AI and Freedom of Expression* and link it to the SAIFE Policy Manual as a tool for both discussion and action in local and regional settings. |

| | | | |
|---|---|---|---|
| **11.15-13.15** | **Assessing and auditing algorithmic content governance: How to ensure meaningful stakeholder engagement?** | **Max Gahntz** (Mozilla), **Claire Pershan** (Mozilla), **Marlena Wisniak** (ECNL) | Audits and human rights impact assessments of AI systems can contribute to a platform ecosystem that better safeguards both people’s safety and their freedoms online (and offline). Recent and ongoing regulatory efforts in Europe, such as the EU’s Digital Services Act (DSA), include obligations for conducting risk assessments and independent audits, with the ambition to define independent auditing requirements and industry standards. This presents a pivotal opportunity to establish an effective and rights-respecting platform auditing ecosystem, and to do so globally. Fundamental questions remain, for instance related to stakeholder engagement, methodology, scope, independence, public disclosure, remedy, quality assurance, and credentialing, to name a few. This session will provide an introduction to audits and impact assessments. Centering meaningful stakeholder engagement as a key component, the speakers will use the example of the DSA to briefly present the status quo of independent platform auditing and an overview of the DSA audit and risk assessment provisions, followed by an open conversation around opportunities for improvement. Drawing on ECNL and Mozilla Foundation’s work on audits, impact assessments, and stakeholder engagement, the session will include an interactive workshop element. The collaborative exercise will look at a hypothetical platform’s content policy for crisis-related information. Participants will explore how audits and impact assessments could be conducted in practice in such a context, with meaningful participation from external stakeholders.<br>This workshop aims to educate policy makers about the emerging fields of (algorithmic) impact assessments, audits, and meaningful stakeholder engagement in regulatory frameworks. It will also provide an opportunity for participants to critique existing approaches to impact assessments and audits, and explore how they could be conducted and regulated for algorithmic content governance systems while also encouraging the growth of an independent public interest auditing ecosystem. |
| **16.15-16.45** | **Bias in offensive speech detection algorithms** | **David Reichel** (EU FRA), **Nienke van der Have** (EU FRA) | Building on its larger body of research on artificial intelligence and fundamental rights, the EU Agency for Fundamental Rights (FRA) offers to present a new report on bias in algorithms. The report includes results of experiments that show how bias in algorithms may arise. One example focuses on bias in speech detection algorithms against certain groups based on ethnic origin and gender. The results contribute to policy discussions around regulating the use of AI and online content moderation. |
| **16.45-17.15** | **Explainability as a precondition for transparent and accountable AI** | **Matthias Kettemann** (University of Innsbruck) | The Explainability Clinic will start with a primer on the legal, technical and design challenges of providing good explanations for AI-based decisions. Explanations are necessary in order to ensure that decisions are accepted and, alternatively, can be questioned. Any decisions that impacts right and obligations needs to be explainable and explained, justfiable and justified. Reasons are essential output, but also the premise of the validity and the precondition of the legitimacy of content-related decisions. We will then proceed with applying what we've learned to three examples from the fields of content moderation, for AI filtering (what's deleted) and AI recommendations (what's amplified) and ad speech. |

## Room 532

| Time | Title | Speakers | Description |
|---|---|---|---|
| **9.00-10.00** | **Regulation digital platforms to secure information as a public good: UNESCO consultations** | **Matthias Kettemann** (University of Innsbruck), **Kateryna Kruk** (Meta), **Andrew Puddephatt** (Sigrid Rausing Trust; online), **Ana Cristina Ruelas Serna** (UNESCO), **Anida Sokol** (Medicentar Sarajevo), **Deniz Wagner** (OSCE RFoM) | UNESCO Social Media 4 Peace project funded by the European Union, has revealed gaps in current policies and practices for managing online content, especially hate speech and harmful content. In many countries, regulatory frameworks are either absent or insufficient or strongly restrictive affecting peoples' right to freedom of expression. In addition, platforms have not invested the resources needed to adequately detect and deal with potentially harmful content online while respecting freedom of expression, particularly at the local level. In this context, UNESCO is piloting a guidance for regulating digital platforms to secure information as a public good to help member States in their regulatory efforts. The guidance will be presented and discussed at a global conference that will take place in February 2023 at UNESCO. Moreover, OSCE has strong research and background work related to digital platform regulation and the promotion of Freedom of Expression and a long-standing cooperation with UNESCO in different areas. Therefore, this session is an opportunity for UNESCO to debate and bring forward ideas to better shape this guidance for regulating digital platforms to secure information as a public good.<br>Therefore, the objectives of this session will be to:<br>• Showcase the testimonies and findings of the Social Media 4 Peace Project, which informs UNESCO's approach to address online harmful content.<br>• Facilitate a debate around the UNESCO approach for shaping digital platforms for the public good, as an open consultation ahead of the global conference in February 2023 and correlate to the worked and experience gather by OSCE members.<br>• As this session consists of a consultation session, the discussion will be led in a participatory manner so that both speakers and audience members will be able to actively interact, and share questions, comments, and inputs. These will subsequently be considered for reviewing the approach to content that damages democracy and human rights to be presented at the conference. |
| **10.00-11.00** | **WARking together: common efforts for better content moderation in Ukraine** | **Tetiana Avdieieva** (DSLU), **Mykola Balaban** (Ukrainian Ministry of Culture and Information), | Effective cooperation between the State, CSOs and social media remains the key solution for most content governance issues, preservation of the evidence of international crimes, combating disinformation, and delivering life-saving information on humanitarian issues.<br>The triangular relations between the state, platforms and civil society serve as a key condition for effective and qualitative content governance in times of armed conflict. The Russian full-scale invasion of Ukraine triggered an intense cooperation mechanism between the various stakeholders, as well as identified a line of outstanding issues to be addressed as a part of the crisis-response mechanism. The Digital Security Lab Ukraine (DSLU) team plans to engage the representatives of social media, Ukrainian government, national CSOs and international human rights |

| | | | |
|---|---|---|---|
| | | **Maksym Dvorovyi** (DSLU), **Marwa Fatafta** (Access Now), **Kateryna Kruk** (Meta), **Marlena Wisniak** (ECNL) | organizations to discuss the cooperation between the relevant stakeholders, extract positive experiences and look for potential solutions to the defined challenges.<br>Intended outcome:<br>• Share practices on cooperation with social media during the war in Ukraine;<br>• Define challenges in the area of content governance, in particular regarding the application of AI-driven tools;<br>• Brainstorm the framework for solving the outlined challenges; and<br>• Build a platform for a broader dialogue with social media platforms on content moderation in times of armed conflict. |
| **11.15-12.15** | **AI and the right to entrusted, secure and encrypted communication** | **Elina Eickstädt** (Chaos Computer Club), **Monica Horten** (Open Rights Group; online), **Daniela Kraus** (Concordia), **Thomas Lohninger** (epicenter. works), **Hedwig Wölfl** (die Möwe) | In his most recent report the High Commissioner of Human Rights of the UN stated the importance of encryption for upholding fundamental rights. However, most recent regulations continue to endanger the right to entrusted, secure and encrypted communication. Additionally, the debate on hate speech and illegal content continues. With technologies such as AI-based detections systems emerging, new considerations need to be made to uphold fundamental right, including freedom of expression and freedom of the media. The impact of AI-based content filters will be a main focus point of this session. In addition, the session will highlight the impact the undermining of end-to-end encryption would have on the right to privacy and existing bans on general monitoring of online communications.<br>The session will focus on AI and the right to entrusted, secure and encrypted communication and will – following an introduction on encryption-weakening technologies and automated decision-making systems detecting illegal content – be interactive by providing and discussing case scenarios where the use of such technologies could exploit vulnerable groups, and the impact this would have on freedom of expression and media freedom. |
| **12.15-13.15** | **Gendered hatred and disinformation: AI – friend or foe?** | **Asha Allen** (CDT; online), **Lucina Di Meco** (ShePersisted; online), **Kyle Matthews** (MIGS), **Lauren Salim** (MIGS), | Around the world, women politicians and journalists report being very concerned that online harms are having a chilling effect on women's political and civic engagement, curbing their ability to speak freely online for fear of threats and abuse. Social media platforms and governments alike are keen to use machine learning-based automated content analysis tools to detect information, and potentially even to delve into using these tools for preventative measures. Machine learning poses an opportunity to counter hate speech through its ability to understand some context of the discussion, how the language is used, the relationship of the speaker to the group or person being harassed and to automate the flagging of content for removal. However, it is not without very serious challenges, including how to build unbiased models, and claims that its use inhibits freedom of speech and produces too many false positives. Does the use of machine learning tools to counter gender-based violence online cause more harm than good? Can we improve our use of AI to counter online harms? What would that look like? |

| | | | |
|---|---|---|---|
| | | **Lorna Woods** (University of Essex) | Using the RFoM SAIFE Policy Manual’s recommendations on the importance of transparency, respecting human rights, providing effective remedy and redress, as well as highlighting the positive use of AI to create safe community-driven spaces, this session aims to discuss the impact of gendered disinformation online, the limitations and opportunities of Automated Decisions Making (ADM) content moderation tools in this context and the importance of an intersectional approach in developing solutions. In addition to the SAIFE Policy Manual, this session will build upon findings from MIGS’ “Canadian Women Leaders’ Digital Defence Initiative" project and CDT’s work, including their report on the impacts of online abuse on women of colour political candidates. |
| **16.15-17.15** | **Blast from the future: Free speech risks posed by emerging tech** | **Keri Lloyd** (Article One), **Michaela Mantegna** (Berkman Klein Center; online), **Mathana** (IEEE), **Eliska Pirkova** (Access Now) | Augmented reality is a broad term for technology that allows us to see our environment with an overlaid digital layer. The objects you see in the real world are enhanced by computer-generated perceptual information combining the real world with a virtual one, allowing real-time interaction and a 3D registration of virtual and real objects. This technology can be deployed using different gadgets, like your smartphone (think of photo filters on Snapchat or Instagram) or with wearables such as glasses. Metaverse firms use Augmented Reality to captivate the attention of customers by allowing them to enjoy the spectacular vision of the physical and digital world. This also helps to increase physical engagement with digital tools, making the users accustomed to new technologies. We often hear that social media platforms have become a de facto public sphere, and how this creates huge problems for the exercise of rights such as freedom of expression. With the proliferation of AR technology especially in public spaces, we risk stumbling into a situation where these same companies control an unregulated “augmented sphere” imposed on our public spaces. There will be content governance issues in AR spaces, with implications for free expression. For instance, mixing deepfakes with AR, something that Walter Pasquarelli calls Synthetic Reality, carries serious risks for groups and individuals. What happens when we are able to virtually place objects — including offensive, harmful, and illegal objects or slogans — on top of real world locations? Will far-right groups use AR to label the houses of immigrants and asylum seekers? Will perpetrators of cyberviolence against women place offensive objects in the “virtual garden” of a victim’s house? These hypothetical scenarios are just a hint of the challenges AR is likely to pose. The panel will map out the risks imposed by AR technology to human rights, especially in the context of freedom of expression and opinion, protection of human dignity and equal treatment. It will provide a starting point for developing human rights centric recommendations on how to build proper safeguards to mitigate these risks, addressed to companies, regulators, and the digital rights community. |

| Ratsaal | | | |
|---|---|---|---|
| **Time** | **Title** | **Speakers** | **Description** |
| **14.00-15.00** | **PANEL I Content Governance in Times of Crises** | **Tetiana Avdieieva** (DigSecLab Ukraine), **Marwa Fatafta** (Access Now), **Julia Haas** (OSCE RFoM), **Matthias Kettemann** (University of Innsbruck), **Marlena Wisniak** (ECNL) | Building on the SAIFE Policy Manual and a workshop held with civil society and academia, this panel aims at increasing understanding of the specific challenges of content governance in crisis situations and at contextualizing existing recommendations on transparency, accountability, inclusiveness, and public oversight. The panel will provide guidance to OSCE participating States on how to ensure free speech safeguards for automated content governance in the context of crises, be it conflict, climate change, or COVID. This guidance aims at contributing to a healthy and vibrant information space in the digital realm – at all times, conducing democracy, sustainable development, and comprehensive security. |
| **15.00-16.00** | **PANEL II Public Interest Framework in the Digital Realm** | **Max Galler** (Facebook Oversight Board), **Hans Hoffmann** (EBU), **Eleonora Mazzoli** (LSE), **Katarzyna Szymielewicz** (Panoptykon), **Deniz Wagner** (OSCE RFoM) | This panel will explore how the challenges in the context of AI-based content governance identified in the SAIFE Policy Manual  can be addressed through a public interest framework and content recommender systems that ensure due prominence to media and public interest information. The panel builds directly on a recommendation provided by the Advisory Group of Eminent Experts on Freedom of the Media for the 25th anniversary of the mandate of the OSCE Representative on Freedom of the Media. The panel will provide a starting point to discuss AI-based opportunities for a healthier digital public sphere, including recommender systems. |
| **17.15-17.30** | **Closing** | **Jürgen Heissel** (OSCE RFoM) | Wrap-up and concluding remarks |

| Room 525 | |
|---|---|
| **9.00-18.00** | Open workspace – can be used by participants for meetings throughout the day |

| Foyer on 5th floor | |
|---|---|
| **11-11.15** | **Coffee Break** – *coffee, tea, and snacks provided* |
| **13.15-14** | **Lunch Break** – *lunch buffet* |
| **16-16.15** | **Coffee Break** – *coffee, tea, and snacks provided* |