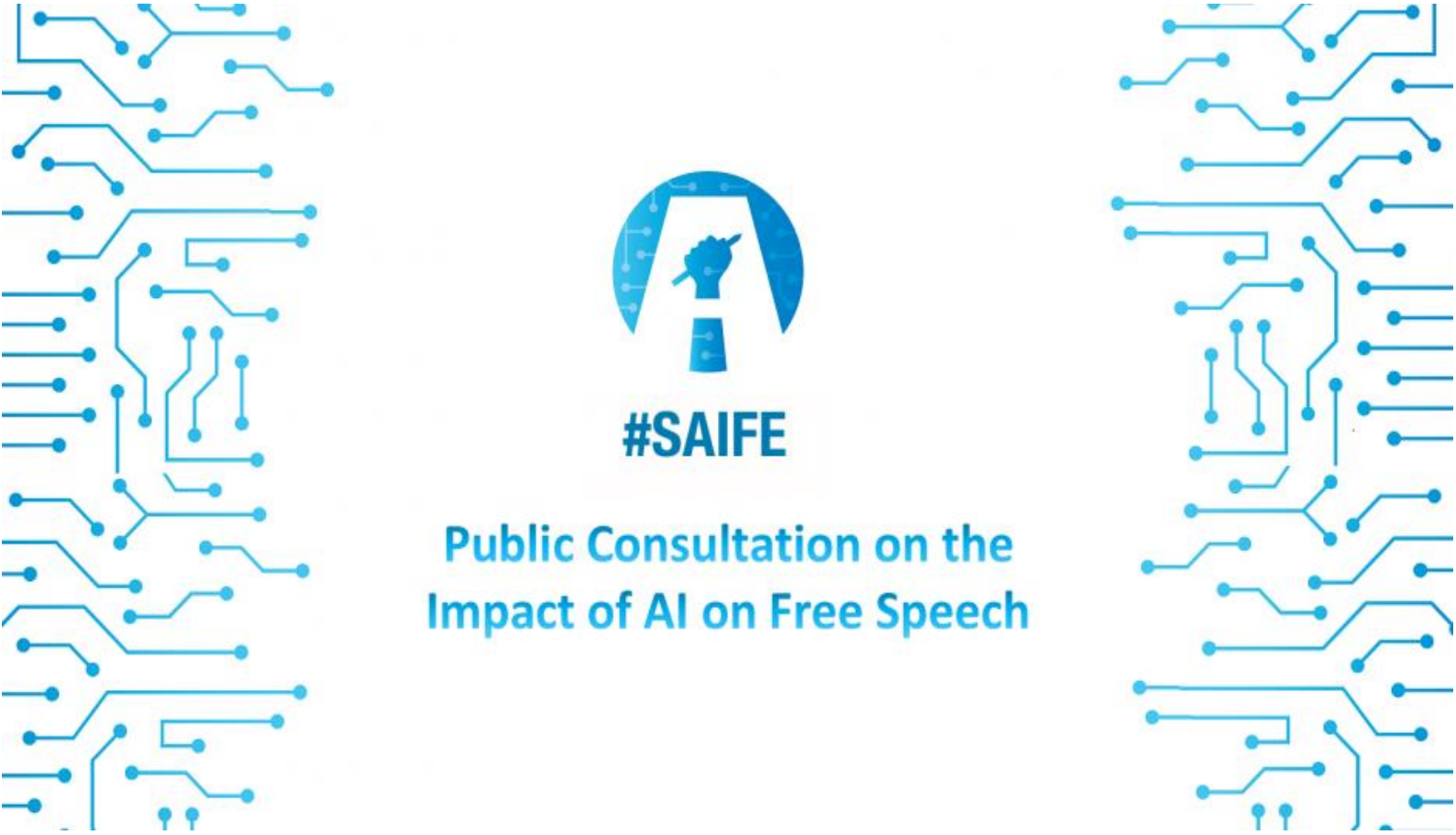
OSCE
The Representative on Freedom of the Media

COMPILATION REPORT
PUBLIC CONSULTATION ON THE IMPACT OF AI ON FREE SPEECH



Introduction

The Office of the OSCE Representative on Freedom of the Media (RFoM) conducted an online public consultation in the framework of its project to put a spotlight on the impact of artificial intelligence (AI) on freedom of expression (#SAIFE). The consultation followed the publication of the #SAIFE Strategy Paper with preliminary recommendations for OSCE participating States and the event on the rise of AI and how it will reshape the future of free speech. The objective of the consultation was to collect a variety of views and opinions, and to obtain input and feedback in view of identifying safeguards for freedom of expression. Providing a comprehensive and inclusive approach, the consultation significantly informs the RFoM Office in developing policy guidelines for OSCE participating States and other relevant stakeholders, including internet intermediaries, on how to safeguard free speech when AI is deployed.

Based on the #SAIFE Strategy Paper, the RFoM Office developed a questionnaire in co-operation with independent experts from civil society and academia. The online consultation consisted of 38 open questions covering the impact of AI on free speech, including specific questions on content moderation and curation, transparency, accountability and the COVID-19 pandemic. The survey was accessible online from July until October 2020, with the possibility to submit contributions until the end of 2020. Alternatively or in addition to the questionnaire, written contributions to the consultation could be submitted.

This compilation report takes stock of the contributions, both through the survey and additional submissions, and presents preliminary trends that emerge from them. Following a summary, the report outlines each of the survey questions with a compilation of the respective responses received. Respondents could contribute to all or some of the questions, and had the possibility to submit reports and other papers. Participants of the questionnaire could fill in the survey anonymously or provide their names (90 per cent of respondents indicated their wish to remain engaged in the #SAIFE project).

The consultation targeted all interested stakeholders and the RFoM Office received responses from representatives of government, academia, civil society, independent experts and the tech industry. Responses came from 17 different countries across the OSCE region (Austria, Belarus, Germany, Italy, Ireland, Malta, the Netherlands, North Macedonia, the Russian Federation, Serbia, Switzerland, Ukraine, the United Kingdom, and the USA) and beyond (Australia, Lebanon).

This report provides an anonymized summary of the contributions. The views presented reflect those of the stakeholders that participated in the public consultation and do not necessarily represent the views of the RFoM Office.

The outcome of this consultation provides useful multi-stakeholder input for the further implementation of the #SAIFE project and will feed into expert meetings in 2021. This input will assist in developing recommendations on how to guarantee the protection of free speech, access to information and media pluralism when AI is deployed. The RFoM Office wishes to thank all participants as well as all OSCE delegations for forwarding the survey to relevant stakeholders.

Summary

The compilation report highlights overarching themes identified during the #SAIFE consultation phase. The survey responses and submitted contributions recognized a variety of challenges to freedom of expression, freedom of opinion and media freedom posed by the use of AI. Many submissions identified AI tools as inherently flawed and their application as problematic. In addition, several solution-oriented and risk mitigating measures had been proposed. The following ten aspects have been identified as particularly relevant for the realization of the right to seek, receive and impart information and ideas of all kinds:

1. **Enhancing transparency**
   The profound opacity regarding the use of AI – including its design, development and deployment over time – can thwart the understanding of policies and practices in place, respondents note. Such understanding, however, would be a precondition to finding appropriate responses to existing and emerging challenges. It is therefore noted as necessary to provide regulatory responses to the use of AI, as it is utilized to enforce private actors' terms of service in opaque and potentially arbitrary ways, which may not necessarily be in line with international human rights standards. A multi-tiered approach to transparency and explainability is suggested, with calls for much needed access to data for independent research. Moreover, in order to increase user agency and empower netizens, respondents underline that transparency in this regard must go hand in hand with increased efforts toward digital literacy.

2. **Addressing context-blindness and profit-oriented content curation**
   A deep concern over the use of AI in content moderation is raised, due to its current inability to understand nuances and context, which may lead to over-blocking and private censorship, including for commercial interest. Moreover, AI-powered content curation tools and revenue-driven algorithmic recommendation systems that prioritize user engagement at the risk of distorting public debate are declared equally concerning. These mechanisms, together with targeted advertising, have been identified as potential root causes of the amplification of discriminatory, divisive and harmful content, as well as for creating and deepening filter bubbles and echo chambers.

3. **Challenging business models and market dominance**
   Surveillance-based business models (including the overall "surveillance capitalism") were identified as an underlying challenge, especially in the context of the gatekeeper function that very few dominant internet intermediaries have in controlling the flow of information and public debate online, including of news content. Such concentration of power, which includes ownership of data and technologies, is stated as particularly worrisome. The so-called lock-in effect – in which users feel bound to certain platforms with a lack of alternative models – is noted to be a result of and to risk stifled innovation and market alternatives. Speech monoculture dominated by a single rule is stated to undermine the value of pluralism, respondents therefore recommend to explore instruments such as interoperability, unbundling of services, or other existing tools from competition and data protection law.

4. **Ensuring data protection**
   In this context, "surveillance capitalism" and severe concerns over lack of data protection were identified as key issues. Privacy is identified as a prerequisite to the exercise of freedom of expression, therefore privacy by design should be considered as a possible solution.

5. **Guaranteeing non-discrimination**

Disproportionate harms on marginalized voices, including through discrimination is noted as an overarching challenge resulting from the use of AI for content moderation and content curation. Over- or under-inclusion in datasets and AI development, as well as discriminatory applications of AI at scale may significantly reinforce existing inequalities and hamper fairness and pluralism. A need to address discriminatory aspects of AI tools and biased datasets has been identified, including through better training of data, diversifying AI workplaces, and increasing participatory processes and human rights literacy of human moderators and those labeling data and AI classifiers. Moreover, multi-stakeholder engagement forums, or decentralized community-based and localized models are recommended.

6. **Judicial responsibilities**

States are mandating private actors to moderate online content, and this is incentivizing the use of automated tools which may impede human rights protection, respondents stress. In this regard, governments instructing private actors with measures that would fall short of their own human rights requirements, e.g. in the context of the use of AI to detect and remove terroristic content, was identified as extra-legal censorship, which risks significantly limiting journalistic work.

7. **Human rights impact assessments and reporting mechanisms**

The lack of due diligence and impact assessments are identified as both a root cause and an amplifying factor to the aforementioned challenges. A priori and regular human rights impact assessments as well as independent audits through the entire AI life cycle are noted as necessary in order to address such challenges. Further, the need for strong reporting tools and transparent and timely complaint mechanisms with options for meaningful redress (internal and/or independently) have been raised as necessary tools for remedy.

8. **Accountability**

The lack of independent oversight and accountability frameworks are identified as significant barriers to preventing and addressed many of the existing impacts of the use of AI on freedom of expression. Self-regulation has proven insufficient in addressing these challenges. Consequently, some suggest a Social Media Council, others draw inspiration from public service media or certification models, and then there are some who call for strong regulatory frameworks that put human rights at their center. Highlighting the deficiencies of one-size-fits-all approaches, many suggest different accountability safeguards depending on the service, the respective actors and the specific human rights risk. In this context, participants note that the use of AI by States and by private actors may require different responses. In any case, there is a call for participatory and evidence-based policymaking. In this context, the UN Guiding Principles on Business and Human Rights is highlighted, as it requires companies to mitigate and remedy human rights harms that arise from their products. While companies have a responsibility to protect human rights, States' must still uphold their positive obligations. Any restriction to freedom of expression should therefore be subject to appropriate oversight, due process, transparency and remedy, based on the conditions of legality, legitimacy, and necessity and proportionality. In this context, the need to foster free, independent and pluralistic media as a cornerstone of democratic public debates has also been emphasized.

9. **COVID-19 pandemic as an exacerbating factor**

The COVID-19 pandemic has been identified as an exacerbating factor, not only because lives increasingly shift online but also due to the increasing use of automation and technology, while human moderation decreased and redress remained limited. Emphasizing that technology is no silver bullet, crisis-specific responses should be considered, while remaining vigilant about emergency legislation.

10. **Ways forward**

Throughout the submissions, participants indicated ways forward for the OSCE. In particular, the OSCE should provide guidance for a variety of stakeholders, building on a multi-stakeholder and holistic approach. It was suggested that the OSCE should endorse the principle of proportionality and assist States in identifying policies to ensure that the design, development and deployment of AI is rooted in human rights. In this context, additional awareness of the use of AI and understanding of the impact on individual's freedom of expression, and the wider societal risk they can carry for freedom of the media was identified as necessary in view of safeguarding free speech and journalists' ability to report freely.

Overview of questions and responses

The text below provides an overview of the 33 thematic questions and summarizes the responses thereto.

1. *The RFoM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (#SAIFE) outlines various challenges to free speech when AI is deployed. What do you consider to be the biggest risk for freedom of expression when it comes to the use of AI? Please specify and, if possible, provide examples on what you, in your field of work, expertise or region, would consider to be the most important issue(s).*

Respondents expressed overall concern regarding the lack of **transparency** of the use of AI, as well as the lack of **awareness** of users and lack of **digital literacy**. Survey participants expressed concerns that this lack of transparency — even about the criteria by which AI filters and prioritizes content — leads to a massive lack of **user agency and oversight**. Respondents emphasized that internet intermediaries' terms of services act as a form of universal legislation, despite being **arbitrarily and opaquely enforced**, which hampers oversight.

Participants also expressed concern that the use of AI leads to **over-blocking**, overbroad **censorship** and **discrimination** as AI lacks understanding of nuances and is often not put into social, historical or regional contexts. It was highlighted that the **outsourcing** of content decisions to private companies shifted control over speech from individuals to companies, and that an ex ante automated moderation represent a de-facto irreversible decision. Additionally, it was emphasized that restrictions on freedom of expression not only stem from **content removals** but also from **recommendation systems** that determine who sees what online.

Moreover, many respondents stressed the lack of **data protection** and insufficient **privacy** as overall concerns of deployment of AI, and that AI is often based on **surveillance** business models. Respondents highlighted that a profit-oriented dissemination of content inevitably focuses on personalized content. Therefore, a technology-fueled media diet might lead to **filter bubbles** and **echo chambers**, and thereby to a loss of **diversity** and variety, and a potential **distortion of reality**. Respondents stressed that the use of AI in content moderation and curation with the underlying goal of maximizing clicks and time spent on services might lead to the promotion of **divisive content**, **amplifying social inequalities** and **attacks against women and marginalized voices**. Today's information overload may lead to users preliminarily consuming information that reflects their own beliefs, participants stated, which may then be learned by AI and **exacerbated at scale**. Respondents were also critical of the potential to shape and modify behaviour through the monetization of behavioural patterns and **micro-targeting**.

The impact of the use of AI on **media outlets** and the lack of algorithmic impact assessment were additional concerns. Potentially biased datasets that lead to AI discrimination, the lack of effective remedies and appeals as well as lack of **access to data** for analyses by researchers and civil society were also highlighted.

Respondents emphasized that the use of AI altogether exacerbates free speech challenges, which increased during the COVID-19 pandemic due to the ever increased deployment of automation and which risk remaining in place after the pandemic.

2. *How can the understanding of the implications of AI on free speech be increased among all stakeholders, including States, internet intermediaries and the general public? Please provide examples of good practices.*

Most respondents emphasized the need for more **transparency of internet intermediaries**, including through regular reports, **human rights impact assessments** and **open data access regimes**. Participants emphasized that comprehensive and comparable **transparency reports** would stir discussions and **inform regulators** on how to address the challenges stemming from the use of AI.

Useful **awareness-raising** and social media **campaigns** could include short stories, examples and citizen projects as well as case studies and research on the fragmentation of the public sphere and diminishing diversity. Overall, it was highlighted that **research** and **studies** should be promoted, as there are still many uncertainties, including on the interaction between engagement and offensiveness of content or the inaccuracy of AI systems removing content. Based on increased awareness, civil society and others, for example re-skilled free speech experts, would be in a better position to put pressure on intermediaries. The **role of the media** (i.e., data journalism) in informing the public was stressed, which should go hand in hand with **media literacy** and **educational resources**.

Respondents underlined the need for a **multi-stakeholder** approach and inclusiveness, as well as **evidence-based policy-making**. Existing multi-stakeholder forums and multi-stakeholder networks, as well as inclusive national debates were mentioned. **Trade associations** for the advertisement sector could also play an important awareness-raising and self-regulatory role.

Survey participants stressed that **users should be empowered to make their own choices**, including on the criteria for the dissemination of content. Overall, comprehensive **legislative frameworks** could increase understanding and provide confidence to users, while avoiding non-compliance. In this context, the **EU General Data Protection Regulation** (GDPR) and its development process were mentioned as good practice examples.

3. *The #SAIFE Paper introduces preliminary recommendations for OSCE participating States and internet intermediaries. For which other stakeholders, if any, should recommendations be developed?*

Respondents outlined the need for **policy recommendations for all relevant stakeholders**, including all branches of States, politicians and parliaments, but also regulators, including for competition. The role of international organizations and co-regulatory media bodies was also mentioned.

Participants noted that recommendations should address public watchdogs and civil society as well as academia and educational institutions. Participants suggested that those who sell and provide AI technology should be addressed, as well as human moderators and partnering organizations, such as fact-checking bodies and those who "flag" content. Respondents underlined that it will also be crucial

to address AI developers, engineers and software developers as well as economic stakeholders and lobbies.

Survey participants also indicated that guidance for the media and journalistic associations should be developed, including regarding their editorial processes. Lastly, respondents stated that end users and content creators should be addressed for better "informational self-determination" as well as the general public with a focus on youth. In this context, the need for a new "netiquette" was underlined.

4. *In the use of AI on free speech. Internet intermediaries, especially social media platforms, act as gatekeepers by engaging in the selection of information that is published, in the ranking and editorial control over it, as well as in the removal of content. For these interventions, AI-powered tools are often deployed. The business model of most internet intermediaries, which is advertisement-based, builds on the collection and processing of massive amounts of data about their users, feeding into these AI-driven tools. Another underlying issue is that a few dominant internet intermediaries act as particularly powerful information gatekeepers in the online ecosystem. Are there additional underlying issues that need to be taken into account? Please explain.*

Respondents identified several underlying issues, similar to those highlighted in the #SAIFE paper. It was emphasized that, in the current internet ecosystem, some intermediaries essentially own the **public space**, for both public and private exchange. Participants expressed concerns regarding this **concentration of power** and **ownership of data and technologies**.

The lock-in dynamics and **lack of alternatives** for users were mentioned as additional concerns, especially due to the lack of interoperability of services and the **information asymmetry** between intermediaries and end users. The extensive data collection and processing to fuel AI mechanisms was mentioned as an underlying concern, and that **micro-targeting** might manipulate behaviour. Moreover, participants expressed concern over the relationship between the use of AI and financial benefits, as advertisements are also placed on "harmful" or even unlawful content, which can result in the trending of extremist content, and on disinformation, which counts as immense profit for intermediaries. In the current ecosystem, counter-speech in the form of re-sharing problematic content with a response might be assessed as increased user engagement and lead to higher popularity of such content.

Another overarching concern was **news blocking** and the financial impact of AI-deploying intermediaries on traditional media. The lack of privacy for end users and lack of effective **complaint mechanisms** was underlined, for those whose content was removed or down-ranked as well as for protection against online harassment by means of AI-powered mechanisms or bots. In this context, **anonymity** of hate speech perpetrators was also mentioned as challenging in some instances.

Participants also expressed concern over the intentional use of **AI for censorship**. Moreover, while different jurisdictions and territorial boundaries pose challenges, it was mentioned that law enforcement and judiciary functions — as well as censorship — is often outsourced by States to private companies, with strict timeframes and high fines. Finally, the need for content policy **standardization**

and global agreement on red lines was emphasized, as well as the need for **professional codes**, where violations have consequences.

Other underlying concerns, such as the lack of **independent auditing**, sometimes **biased datasets**, and AI tools not being built in an ethical or diverse manner, were mentioned, and that the use of AI might reinforce biases — especially without a human in the loop, and with content dissemination at scale. The risk of over-blocking, limitations of current AI tools and their inaccuracy due to the lack of understanding of **linguistic diversity** and equity were also underlined.

Overall, respondents agreed that **self-regulation** has proven to be insufficient in addressing these challenges, especially as all content, from copyright infringements to terrorist content or hate speech, are treated the same way despite requiring very different responses.

5. *The #SAIFE Paper addresses the role of "surveillance capitalism" business models in creating AI systems that can threaten freedom of expression and privacy. What alternative business models, besides behaviorally targeted advertising and the monetization of users' data, can support hosting of user-generated content on a massive scale? Which alternative business model could better protect diversity of voices online, and how could such a model be designed for different kinds of content-hosting and other online services? Please provide examples of good practices.*

Survey participants underlined that the **internet as a public space** should neither be privatized nor monopolized. Alternatives should be available and based on strong **data protection** and **competition** laws, which are strictly enforced, as self-regulation has proven to be insufficient. Respondents proposed different alternatives, such as **fragmenting dominant internet intermediaries' services**, as their **gatekeeper** function stems from the fact that user-generated content is not only hosted but also moderated and curated by the same actor. Opt-in defaults could encourage more **competitive models** with functional separations. Other proposed concepts included so-called "freemium" **subscriptions**, where the basic service is free but more advanced features are fee-based.

Respondents noted that alternatives should be **decentralized** and **community-based models**, involving non-profit services on a local level. Interoperability could be a solution, as could specific bans on the monetization of data without prior consent, on the individualization of political messages through micro-targeting or overall prediction through surveillance-based models.

Others proposed **public broadcasting**-style platforms with significant funding and support as services for public interest, noting that essential services are provided by non-market mechanisms, such as roads, postal services, electricity or water. One respondent referred to the need for **digital taxes** so that alternative social networks and providers can grow; while another proposed "social impact bonds" to **fund diversity initiatives**. Growing alternatives should address the current **digital divide** and not be allowed to be bought up immediately. Additionally, it was mentioned that users should be incentivized to develop their critical thinking.

6. *Do you consider the dominance of a few internet intermediaries to be a challenge for freedom of expression online? If so, which specific concerns do you see?*

The main challenges identified by participants on the **market dominance** of a few companies were that these companies are global and the sole **arbitrators of speech**, where few private actors dictate terms of services, policies and practices determining speech. Respondents stressed that private conditions need to be accepted to access information and participate in public debates, which makes these oligopolies **economic and human rights gatekeepers**, which is strengthened through **data control** and **advertising**. The lock- in and network effects reinforce this situation. While terms of services dictate human rights decisions, respondents underlined that they can be changed opaquely without prior notification of the users, even through government pressure. Overall, survey participants noted it to be particularly problematic if **public services** are embedded in those dominant intermediaries.

At the same time, the incoherent implementation of platforms' terms of services without **remedy or redress** was emphasized as a challenge. Companies could manipulate information flows and shape what knowledge is available and acceptable, for example when it comes to different approaches towards free speech and nudity. It was also emphasized that global market power leads to **culture bias** through the global application of terms of services, where local aspects are disregarded. Terms of services might be developed independently from **human rights** and **international law**, and might therefore be too restrictive or loose, while their enforcement often remains opaque. At the same time, it was noted that States outsource decisions on what is acceptable speech to private actors away from courts pursuant to law.

Respondents also underlined that **market dominance** leads to less diversity, which equals lower quality content. At the same time, it was mentioned, market dominance leads to a lack of incentives to create and share content for and by **marginalized groups**, or indeed to prioritize content of public interest or accurate information. Moreover, users can be easily used as "guinea pigs" of dominant internet intermediaries as they reportedly deploy different algorithmic systems to different users in order to identify the most profitable content curation mechanism.

In addition, some respondents underlined that **traditional media** are suppliers and competitors of intermediaries at the same time, but with an unfair share value and no bargaining power, which, in turn, reinforces questions of **editorial independence**.

Moreover, it was noted that market power influences innovation, incentives for change and constitutes entry barriers for competitors. It was highlighted that democracies should always strive for a **balance of power.**

One respondent outlined that, (much like the financial markets in 2018) the benefits of AI and AI-powered content moderation and curation are privatized, while the costs are transferred to society. Overall, respondents expressed disappointment that the **internet's promise to be an open and accessible communication tool for everyone** has not materialized but instead users are disempowered to increase intermediaries' profit.

7. *Is there a need to create a policy and normative environment that is conducive to a diverse, pluralistic information environment in the AI domain, or to ensure competition to prevent the*

concentration of AI expertise? Should the "network effect"[1] and limited interoperability of online services be addressed? If so, how?

Survey participants emphasized **interoperability** as a response to address the network effect. The EU standards on USB plugs were mentioned as a best practice in this regard. Moreover, interoperability would provide individuals with better control over their personal data and decrease entry barriers, incentivizing competition. It was also mentioned that interoperability can fill gaps in services. Another proposal to address network effects was **open source codes**. Respondents emphasized that independent third party control mechanisms are crucial and can increase good governance.

Moreover, participants underlined the need for addressing **local contexts**, as a network effect paired with disregard of local context can have detrimental consequences. It was stated that global companies have a responsibility to adjust their content moderation to local contexts. Survey participants stressed that addressing the network effect requires enhanced co-operation and finding solutions for jurisdictional disputes.

There seemed to be convergences that if the **network effect** and **economies of scale** enable monopolies, then there is a need for **regulation**. At the same time, however, it was also stated that regulation could entail own challenges, in particular if adopted in a hasty manner without multi-stakeholder processes. Currently, it was noted, some companies have an almost normative power and their policing affects how societies around the globe behave. Therefore, in-depth merger review and competition enforcement would be needed. Regulation could be considered to create "free speech spaces", with internationally recognized rules for user engagement and clear consequences for abusive behaviour. Respondents widely recognized that **the internet is essential** for communication and the exchange of goods and services, just as roads, railways, waterways, airspace or telecom and postal services are – all of which are subject to certain obligations regarding performance, access and content policies.

8. *Is there a need for different approaches or different free speech safeguards on AI depending on the specific internet intermediary, their size, capability, extent of risks of human rights impact, and services offered?*

Participants underlined that while there is **no one-size-fits-all approach**, the principles for freedom of expression are the same for all. Various respondents agreed that the level of necessary **safeguards** depends on the kind of service, however, not necessarily on its scale. Others called for different approaches depending on the number of users and market power of the service provider in question. This could affect, for example, the frequency of **mandatory audits**, **transparency reporting** and **algorithmic auditing**, the need for audits regarding the impact on vulnerable groups, and overall more stringent control for those with wider reach. It could even require an **unbundling of services**, as suggested by a respondent, due to the asymmetric remedy for those with gatekeeping positions.

[1] The network effect is a phenomenon whereby increased numbers of people or participants improve the value of a good or service. A social media platform might therefore grow in popularity because it has achieved a critical mass of users and new users will be deterred from using another platform. For more information, see the #SAIFE Strategy Paper.

Participants noted that safeguards should not include undue financial or resource-intensive requirements on smaller platforms to **avoid anti-competitive effects**. Others proposed to refer to the extent of human rights risks, with a **diversified approach to liability** depending on the service offered. Another respondent emphasized the need to adapt a positive approach to promote "good discourse". All seemed to agree on the need for proportionate approaches in order to reflect international law and human rights standards. Some called for caution as research is necessary before regulatory responses are taken.

9. *The #SAIFE Paper addresses how the surveillance of individuals' activities through AI technologies, by States (often relying on data collected through, and shared by, private companies) and by the private sector resulting from their business model, can seriously impede freedom of expression. What are the main risks stemming from the use of AI for surveillance techniques?*

Respondents underlined that **surveillance** always brings dramatic chilling effects, and even the mere feeling of being surveilled can lead to **self-censorship**, a **change in behaviour** and/or psychological disorders. Survey participants emphasized that mass surveillance is inherently disproportionate, and targeted surveillance should be based on the **three-part test in line with Article 19** of the International Covenant on Civil and Political Rights (principles of legality, legitimacy, and necessity and proportionality). It was stated that the pervasive and invisible nature of AI – combined with its ability to identify and track behaviour – risks chilling speech.

Respondents also noted that the training of AI on historical data risks perpetuating historical problems, e.g., embedding societal stereotyping in data, which might lead to **discrimination.** For example, slang and minority words are **underrepresented** in data, which results in higher error rates. At the same time, participants emphasized that flaws in AI systems might remain undetected due to their **opaque deployment**, so **biases are unprovable** and redress precluded. Respondents underlined that **redress is crucial**, especially concerning AI-powered nudging techniques and the information asymmetry, which leads to an increase in structural inequality. It was stated that COVID-19 increased these risks.

The lack of **explainability** of AI was also mentioned, as it hampers scrutiny and investigation. More research seems to be crucial, especially on how online behaviour influences offline behaviour. It was noted that the speed and scale of AI tools are not understandable to users, and that surveillance models can **distort the information environment** and **intervene with public debate**.

Other concerns raised referred to the risk of storage and leaks of huge amounts of data, as well as the sharing of data with third parties, including States, and questions about control over the data. It was noted that **private surveillance** drastically increases the scope and scale of **State surveillance**. Data and AI techniques could be sold to States and used to interfere with or even prevent protests or alike, which poses a risk to the right to freedom of peaceful assembly, freedom of expression, as well as the safety of journalists, human rights defenders and whistle-blowers.

10. *Can you provide examples of AI-powered surveillance technologies that are used in accordance with human rights principles of lawfulness, legitimacy, necessity and proportionality, and of available legal redress mechanisms for victims of surveillance-related abuses?*

Respondents emphasized that there should be no general AI-powered surveillance, but rather clear **limitations on AI surveillance**. It was noted that, if human rights-based surveillance is at all possible, then it should be done only in niches and with clear transparent, independent and accountable mechanisms. **Social media councils** were suggested as such **accountability mechanisms**. Respondents noted that the use of AI's superior capacity to collect and process data to manipulate behaviour is no human rights-friendly AI deployment. It was noted, once again, that AI is brittle, error-prone and often biased.

11. *How can users' agency and choice regarding the application of AI processes be enhanced to ensure better free speech protection? What role does the principle of "privacy by design" play, or opt-in and opt-out provisions in respect of AI systems?*

Respondents underlined the need to increase user agency and their regaining of bargaining power. It was underlined that viable alternatives of services are needed to achieve this. Also, the **use of AI should be disclosed** just as the use of cookies has to be disclosed in the EU. At the same time, it was noted that not the entire AI code should be disclosed in order to prevent misuse.

Respondents noted that the default interfaces should not use personal data but instead be free speech-friendly, enable interoperability and unbundle services functions. It was additionally stated that **users should be informed about the use of their data** for advertising or for the prevention of illegal activities, so that they can consent to their data being used for different purposes only, or for sharing with specific stakeholders only. Participants emphasized that the standardized protocol should turn off **profiling** for recommendation systems for content, so that any filter is up to the user's discretion and they have the choice over ranking and what information they receive. **Privacy by design** models should be implemented with **media diversity as the default**.

While some proposed opt-in or -out models for personalization and micro-targeting, others suggested more detailed option models to establish a **right to informational self-determination**, suggesting that users should be able to restrict permission on a case-by-case basis.

Proposals to increase user agency included labels, certifications or classifications for safe and privacy-preserving AI applications, with **certificates based on transparent auditing by independent actors**. It was also stated that data should not just be collected because it is technically possible.

12. *The #SAIFE Paper outlines how the use of AI for content moderation can lead to the removal of legitimate expression, or failure to remove content that could have a negative impact on those who access it ("false positives" and "false negatives"). Do safeguards need to be introduced in AI-powered tools to address this? If so, what kind/which ones?*

Survey participants highlighted that **transparency**, including on which tools are involved in decision-making, due process and public oversight are crucial for **accountability**. Transparency reports should include information on whether **laws or terms of services** where the basis for a removal. Moreover, respondents stated, information on the **accuracy** of AI-powered removals, including verifications for them, and **corrective mechanisms** against **error rates** should be made available, as well as

engagement with State authorities. Respondents stated that transparency reports should include information on **appeals** and their outcomes, as well as details on the use of automation, including training data and the criteria used for decision making.

Internal complaint mechanisms were mentioned as crucial, with clear and easy-to-find information and with the possibility to reinstall content during the appeal. As a safeguard, there should be a right to apply twice if content is removed. Respondents underlined that any decisions should be subject to local advisers and adapted to local contexts. There should be possibilities to complain cost-free to independent and impartial bodies, potentially **independent cyber judiciary** and **dispute resolution**.

Independent self-regulation mechanisms were mentioned as useful, as was **collaboration with stakeholders**, e.g., social media councils inspired by journalistic self-regulation. Some respondents stated that AI should only be used for limited parts of moderation due to its inaccuracy. Others stated that any removal and redress should include a human in the loop, with clear information for users.

It was also stressed that research on the effectiveness of removals should be encouraged. This could be done through **open source** methods or by making **independent periodic audits** with public results by civil society and academia mandatory. Removed content should also be made available to researchers to scrutinize them for biases. One respondent proposed "**media colliders**" where AI distils and tones down content instead of removing it.

13. *The #SAIFE Paper outlines how the assessment of the (il)legality of content is a complex task, and depends on local context, local languages, and other societal, political, historical and cultural nuances. AI-driven decisions for content removal can fail to understand nuances underpinning the pieces of content, resulting in the filtering and taking down of legitimate content. Is there a need for a "human in the loop" in AI applications? If so, what level of human review or genuine human involvement should be ensured?*

Respondents expressed particular concern over the **use of AI across cultures and contexts**, without taking into account that political, cultural, economic, social, power dynamics shape users' expression and expectations. It was emphasized that no datasets can accurately account for the fluidity and variance in human language and expression. Hence, participants underlined that AI is often inaccurate, especially as most terms such as "hate speech" or extremism involve fluid definitions. In addition, it was noted that humans learn how to trick algorithms. Therefore, it was highlighted, local moderation and consultation is crucial.

Survey participants noted that the level of **human review** could depend on a platform's size and type of activity. One respondent proposed three levels for human review: **local moderation**, **tech moderation** (in order to amend AI systems where necessary) and **supervisors** (in order to check the AI's viability). Some suggested that human control is needed for cases where the content is not clear, and that humans must test the underlying data, the performance of AI and equality guardrails for AI's deployment and performance. **Quality control** could be enabled through randomized control undertaken manually (internally as well as independently). Again, transparency and understanding of how AI codes work were mentioned as crucial.

14. *Is there a need for different levels of human review depending on the context in which AI is used (e.g., for the curation and prioritization of media content) compared to the use of AI to identify and flag potentially illegal content?*

Different levels of human review could reflect **proportionality**, stated respondents, so that a higher level of review would be needed for opaque or risky human rights decisions. It was stated that the identification of illegal content is a **social goal**, while the curation and prioritization of content follows the purpose of achieving **business goals**. Therefore, they require different approaches. Moreover, respondents highlighted that assessing dubious content requires different qualification standards than illegality.

Respondents emphasized the need for **oversight** over **targeted advertisement** and taking into account local contexts. Given the amount of data available, it was stated, not all decisions can be subject to human review, but oversight is needed over the aggregated impact of prioritization decisions and the design of algorithms. In order to better determine whether offensive or discriminatory content is recommended or flagged by AI tools, respondents highlighted the need to **access data**. A recent study showed that nudity leads to more reach on social media, but there is no data yet regarding offensive content.

15. *What measures should be taken (and by whom) to ensure that societal inequalities in the production of, and access to, information, which impact freedom of expression, are not reinforced in the development and deployment of AI technologies?*

Respondents underlined the need for data that reflects the diversity of the world population. For this, diverse teams are crucial as is including sets of **criteria for diversity** and to evaluate standards. Moreover, respondents mentioned the crucial importance of taking into account local peculiarities in the development of AI and adjusting it regularly along with **engaging local moderation**. There should be co-operation with co-regulatory bodies, and civil society and/or independent experts could recommend adjustments for local content. Moreover, the need for **transparency** over AI models and data used, **independent auditors** and **user empowerment** were mentioned as crucial.

Some respondents outlined the need for a **regulatory approach to ensure diverse datasets** and to **prevent bias re-installment** through technology. This would require a global approach and should include **State indicators** and **transparency obligations**, which could be tiered, so that only trusted third parties receive full access. State and inter-State systemic involvement was underlined as important.

According to some respondents, **tech-facilitated discrimination** could be avoided through **regional audits** and review of the development, outcome and regular use of AI. Such audits should include marginalized groups. Moreover, competition regulation could provide an interesting approach to ensure more diversity.

Survey participants underlined the need for a **multi-stakeholder** and **multi-disciplinary** approach to agree on how to **test methodologies** and to exchange **best practices** to reduce and mitigate bias. In

addition, one respondent underlined that data should not belong to a few internet intermediaries, but rather that **data should be seen as a public good** and be redistributed, which requires a remuneration for the data – both on the individual and collective level.

16. *The #SAIFE Paper outlines how the use of AI can impede the free flow of information and democratic discourse. AI-powered tools are often used to categorize users to determine their particular political, commercial and other preferences in order to target them with specifically curtailed content. What measures could be initiated, by State and non-State actors, to promote the use of AI to foster diversity and to create an enabling environment for media pluralism online?*

Respondents noted that, in order to prevent so-called **echo chambers**, data-based curtailing of content should require **prior consent** and users should have the possibility to change and control recommendation algorithms. For example, it should be possible to choose "**public value**" information on one's newsfeed, and it should be possible to have alternative and local specific tools with a higher representation of otherwise underrepresented groups and topics. Respondents suggested that on sensitive topics, targeting with curtailed content should not be possible, just as curtailing information during certain periods, such as **pre-election periods** or **during conflicts**, and information of public interest should be prioritized.

Some survey participants highlighted that **co-regulation** in special bodies where States and internet intermediaries co-operate could oversee such initiatives. Others proposed bodies where States and civil society join forces to develop rules for internet intermediaries. According to some, international bodies could assess which content should never be subject to targeting. Alternatively, a positive approach could be considered, where constructive debates with **inclusiveness** and consensus-seeking are rewarded.

Various respondents emphasized the need for regulation reflecting **cultural differences**, for example in the case of nudity in art. Legislation for **public service media** could be an inspiration. Another possible tool mentioned was taxation, to stimulate the development of **new players** with different recommendation systems. The need to **support small and regional media** was also mentioned.

Again, **transparency** was mentioned as a crucial pre-condition, including for users to be aware of their categorization and on which criteria they see which information and what the effect is. Respondents noted that it should be clear to users whether information is shown through contributions of AI and whether the information is organic or advertised content. **Accessibility**, **inclusiveness** and **open source** should be promoted.

17. *Should human rights impact assessments for AI-powered tools be mandatory, and if so, how, on which level (design, development, use of training datasets, deployment of AI), and according to which timeframes? Should specific AI applications require specific evaluation prior to their deployment? Who should conduct such assessments, and what should be the response to any foreseeable risks?*

Respondents underlined the need for periodic **human rights impact assessments** (HRIAs), **due diligence** and **continuous auditing**. These should be mandatory and tailored to the specific context

and intended use of AI-powered tools. HRIAs should be comprehensive. Participants emphasized that HRIAs should cover the entire process and full AI life cycle to ensure that **design, development and deployment of AI is rooted in human rights**, including the code making and design and formation of datasets. All international and national efforts should be informed by human rights.

Respondents stated that the frequency of HRIAs could depend on the size of the service and the purpose of the use of AI, while others proposed an evaluation twice a year as a benchmark. Additional review could be required if the collection, assessment and processing of personal data is involved or if data is shared with third parties.

Participants noted that HRIAs need to be conducted by an independent and accountable body that works transparently, while non-disclosure agreements could be used to enable the audit of specific codes. Standards for HRIAs should be regularly updated and the first assessment should be conducted before AI is deployed, with the deployment being abandoned if the risk is too high.

It was also mentioned that HRIAs should not be a barrier to entry into the market and should therefore be easy and affordable. Throughout HRIA processes, respondents underlined, data protection must be ensured.

18. *What measures, if any, need to be implemented to ensure effective remedies for AI-powered tools? How can it be ensured that those impacted by partially or fully automated decisions enjoy protection against erroneous or discriminatory outcomes? What internal complaint and redress mechanisms need to be installed for users in relation to AI?*

For **appeal processes** to be meaningful, respondents stressed, an adequate notice to the impacted user is crucial. Such a notice should take place if content is removed or if an account is banned and should include information on the specific content and terms of service clause that led to the removal/ban. In addition, participants noted that users should be notified about the down-ranking or recommendation of their content. Information should also be provided on how the content was detected (privacy compliant) and removed, as well as on appeal possibilities. Moreover, respondents emphasized, information should include to what extent AI was used.

Respondents stated that appeals should include **human review** to avoid situations where AI assesses problems caused by AI. Users should have the opportunity to present additional information and be notified of the result of the review with a clear explanation. One respondent underlined that **judiciary** and **legal professionals** should be involved in this process. Participants stressed that **appeals** should not include any costs for users, be **accessible** and **timely**, and that the **user's jurisdiction** should be applicable. Not providing appeal mechanisms should involve **strict liabilities** with financial penalties.

Moreover, it was suggested that flaggers of content should have the possibility to **examine a log of previously flagged content** and the outcome of moderation processes. Some respondents proposed that notice and appeal should also be available for flaggers.

It was also stated that there needs to be **independent oversight** as intermediaries' rules are quasi-normative, so internal remedies should be audited and a **third party complaint mechanism** made

available. In any case, participants again highlighted the need for transparency, regarding both which speech is undesired, so users are better informed, and regarding the complaint mechanisms. There should be **transparency regarding appeals**, participants stressed, including the amount of content affected, the process and their success rates. Current transparency reports have a narrow remit and only include information on removals.

Some respondents emphasized that images, video content and other specific types might require **more human review**. Furthermore, it was mentioned that **removals should be temporary** with the possibility to restore content, or content should rather be **prevented from trending instead of blocked**.

19. *The #SAIFE Paper calls for stronger transparency of AI-powered tools in content moderation and curation. What measures are needed to increase **transparency** while ensuring a strong data protection framework? Please provide examples of good practices.*

As in various previous responses, **transparency** was mentioned as a crucial pre-condition for any informed discussion on AI, its impact and the potential need for redress. Additional responses focused on the need for a possibility to request more information regarding how decisions were concluded and that transparency should also include general public information on the AI's functioning.

Data protection was mentioned as being essential in any transparency mechanism, while access to classified information of users should only be provided with users' approval. It was also stated that **anonymity** has a high value but should not be sacred if needed for due diligence and rule of law.

Good practices mentioned were **privacy by design**, where users can control the sharing of their data, and **easy-to-navigate opt-in and opt-out systems**. Participants suggested that **courts** should have the possibility to request additional audits, and any **government-used software** should be **open source**. Advertisement archives were mentioned as a good practice example, specifically as inspiration for how to disclose information on the prioritization of content.

20. *What minimum standards of transparency should be introduced for the use of AI? What elements should these standards contain?*

Respondents listed various aspects that should be covered by **transparency reports**. Information should be provided on a regular basis, in an explainable way and in a structured format, with static links, and should be easy to find and understand. Respondents suggested that higher risk AI or data use could require more transparency. The extent of information can vary, so that users, authorities and researchers receive a different level of data.

It was noted that if **terms of services** are changed, users should be notified beforehand. If special rules are applicable, e.g., during elections or emergency related rules, these should be easy to find and access. Participants highlighted that information should be provided, in clear policies, on which speech is not permitted and how such rules are enforced, including on the process of how breaches are identified. Transparency reports should contain moderation rules, filtering rules and content evaluation and ranking, and hence **all editorial decisions** (comparable with media ownership

disclosure). Information should also be provided on how ranking works and on control options for users. Information should be provided on whether data is disclosed and used for AI (by default or with consent), and the methodology on how the impact of AI is assessed. Information should also be provided on positive attitudes towards inclusion.

Transparency reports should include which content was taken down, based on which rule, how it was detected, if it was appealed against and, if so, what the result was and whether the content was restored – and, if so, whether it was done so proactively because of an error or because of an appeal. Information should also be provided on the use of AI, their accuracy rates and the extent of humans in the loop, as well as on the datasets, what output models they generate and what was done against the misuse of AI models and mitigation measures. Information on due diligence, human rights impact assessments etc. should also be provided.

Participants stressed that information on **data collection** and their purposes should be provided. This should also include information on the associated risk and how the data influences what users see and should include the option that **users can delete their data**. Transparency reports could also include which data sources were drawn on personalization and to which extent. All information on the **use of data for targeting** should be provided, while the **use of data for protective aims** could include general information only. Public databases of advertisements were mentioned as a good practice example and could give inspiration for databases on categories of users and AI-fueled content dissemination.

In addition, it was stated, information should be provided on **public private-partnerships**, in particular on law enforcement requests to take down content. Additional transparency was suggested for paid content and advertisement targeting, and on how many advertisements users see on average per minute, which might require special human review for sensitive advertisements.

21. *Should a multi-tiered approach to transparency be introduced? If so, what should be the main considerations?*

While one respondent negated this question, most seem to be in favour of a **tiered approach** where the amount of available information depends on the recipient of the information (consumer, user, regulator, researcher etc.) and the level of human rights risks stemming from the use of the specific AI tool. This would mean that all intermediaries should be bound by transparency reporting obligations, with certain service providers being subject to additional transparency requirements. Government institutions and international organizations should hence receive **more detailed and technical information**, as well as other neutral and trusted third parties, in particular independent researchers. Thereby, respondents emphasized, the argument of not sharing "**business secrets**" could be avoided. Participants noted that citizens need more information regarding their media diet; media and researchers need information for investigative and research purposes; and the public needs to know the policies and practices that are in place to enable democratic debate and legitimacy over decisions.

Moreover, participants suggested that large-scale services with a significant impact on the public debate should be obligated to disclose more information than small actors. Regarding the disclosure

of user data, respondents noted that the level of threat should be decisive; incitement of hatred should require different disclosure than IP infringements.

22. *How often should transparency reports be made publicly available? Should there be any other criteria for publication?*

Survey participants stressed that **transparency reports** should be published regularly. Respondents suggested that the frequency could depend on the size of the service, the type of AI action and the market position of the one deploying AI – while those with a gatekeeper function should publish reports every three months, small- and medium-sized companies publishing reports annually would suffice. It was also noted that frequency could depend on the AI system and how frequently it is updated as well as the degree of influence on human rights and whether there are proper accountability and oversight mechanisms in place.

Other respondents noted that **transparency reports could be published by independent third parties**, where they access data and decide on the level of public disclosure. Mechanisms to ensure user privacy were emphasized as crucial.

Respondents underlined that transparency reports should be published in a structural format rather than a PDF to simplify data extraction and increase comparability. All transparency reports should be accessible at the same location with static and functioning links.

23. *How can transparency norms and expectations be harmonized to ensure that disclosures are comparable, accurate, and useful to a broad range of stakeholders?*

Respondents stressed the need for **standardization** to enable comparability and better understanding of the overall content moderation and curation landscape. There should be **minimum standards on a global level** and such general norms could be developed by international organizations. It was suggested that specific norms could be developed in a decentralized, bottom-up and multi-stakeholder approach. For this, **certification schemes for AI tech** were mentioned as good practice examples.

At the same time, respondent also referred to the challenges arising from the existence of different free speech regulations around the world. Lessons could be learned from other fields, such as rules concerning sustainability. It was also mentioned that specific metrics could reflect specific content, roles and services, as there needs to be a balance between standardization and specifications for different contexts and services.

24. *Are there any free speech impediments to mandatory transparency regimes that need to be addressed, and how could they be mitigated?*

Respondents referred to the need for special protection of the **anonymity of users** as well as of **minority groups** in particular. Confidentiality and data protection in transparency reports were stressed as crucial. Particular challenges identified by respondents could arise in non-liberal contexts

and due to the lack of consensus on underlying definitions. It was also mentioned that transparency potentially opens the **space for abuse**, which is why, for example, investigative techniques of law enforcement are not disclosed to avoid that they are misused to circumvent the law.

25. *Are different levels of transparency and accountability required for the use of AI in different stages of content moderation and content curation (from uploads, to making certain content more visible for users, to the removal of content), in search results, or for tackling "inauthentic behavior"?*

Respondents emphasized that the principles are the same for all stages and applications of AI, but that the details are different. Different requirements could be linked to the level of threat to free speech and privacy. Survey participants suggested that, depending on the potential harm of AI tools, scale of the service and volume of content, different responses could be adequate, just as administrative justice is not equal to judicial proceedings and norms associated with them. Different approaches may be relevant, depending on the level of user engagement and human involvement and intervention. Others emphasized that the level of transparency should be the same for all stages of AI, but involve different criteria. Information needs to be adjusted as, for example, removals are easier at comparing and identifying errors than biases in search results.

Respondents stressed that a chain of transparency is necessary to ensure trust. In order to avoid misuse and circumvention, the specific code of removals should not be disclosed. At the same time, participants emphasized that rigid information on paid content is needed (including in proportion to unpaid content and its reach) and which data is used as well as how it affects recommendation systems.

26. *Based on increased transparency, where and how do you see possibilities and benefits of multi-stakeholder contributions?*

Respondents stressed that a **multi-stakeholder approach** is essential to ensure a balancing of interest and the finding and filling of gaps in proposed regulation, as well as for the sharing of good practices. Overall, multi-stakeholder contributions from government, civil society, the tech sector and human rights experts on both the national/domestic and international level were mentioned as crucial to assess risks from different perspectives.

Survey participants underlined that multi-stakeholder discussions enable addressing possible abuses of AI technologies and can contribute to media literacy, awareness raising and are important for institutional mechanisms and oversight. The need for a local focus was highlighted.

At the same time, it was also mentioned that including too many voices in regulatory discussions might lengthen any process and lower their quality if too many conflicting values or the promotion of private interests is endorsed. It was highlighted that the current situation, where a few have near-complete power to shape online discourse, needs to be changed. To enable a meaningful multi-stakeholder approach, the willingness of those powerful actors is necessary.

Respondents suggested that specific indicators for multi-stakeholder involvement could be monitored and benchmarked by independent bodies. Again, transparency was mentioned as a pre-condition as it would benefit research, civil society, law enforcement and policymaking.

27. *The #SAIFE Paper emphasizes the need for more accountability of AI-driven tools in content moderation and curation. What measures are necessary to ensure that AI systems employ a variety of controls, to verify that they work in accordance with their intentions, and to ensure that the operator can identify and rectify harmful outcomes or reproducing inequalities? What governance arrangements would lead to an effective system for supervising and ensuring that free speech is protected when AI is used?*

Respondents emphasized the need for an **impartial and independent review system** with pre-defined acceptable error rates. It was proposed that terms of operations for independent advisory bodies with a legal mandate could be developed by international organizations, stressing that any independent authority requires tech expertise. It was highlighted that the purpose and intention of AI tools should be reviewed. To ensure accountability, a clear allocation of responsibilities within audited companies for all processes was mentioned as crucial.

Survey participants stated that both testing of new/updated AI tools and periodic human review before their implementation should be required. Additionally, it was suggested that, where AI is trained to react to unpredicted circumstances and adjust itself, updates should require prior review by humans.

The need to **translate international standards and safeguards into AI tech** was highlighted, referring to data protection as a good practice example.

A **Social Media Council** as an appeal mechanism was proposed as an interesting option. Such a council would enable a multi-stakeholder accountability mechanism as a transparent and independent forum to address content moderation. It could be a voluntary compliance approach (compared to legal obligations) and have both an advisory and adjudicatory role to provide general guidance on international standards as well as the power to review individual content moderation decisions. Such an accountability mechanism would need to be accessible to all and could provide co-regulation in addition to an overseeing public authority. It could be more effective than harsh legislation with disproportionate sanctions and restoring of trust, and easier to adapt to the constant evolution of tech.

28. *What good governance and accountability processes can serve as a model for algorithmic and AI accountability, for all actors involved and at all stages in the process (design, use of training datasets, human rights impact assessments etc.)? Please provide examples of good practices.*

While its effect for competition and user protection remains to be evaluated, the **EU General Data Protection Regulation** (GDPR) was mentioned as a good practice, both content-wise and regarding its inclusive development process. Additionally, **certification models** could provide inspiration as it would include clear criteria for all stages, and the process would enable a better understanding of risks and remedies. Moreover, **openness and transparency** were again mentioned as crucial for good

governance. **International Standards on Auditing** (ISA) could provide inspiration for risk management of software design. Respondents stressed that there should be liability for harmful AI deployment; some respondents also suggested secondary liability on individuals, e.g., humans could be held responsible when testing AI in the working process.

Respondents assessed the ongoing drafting of legal instruments for AI tools as an important approach.

29. *What could and should be the role of independent oversight and of an auditing mechanism to ensure meaningful accountability of AI systems?*

Oversight is only meaningful, respondents emphasized, if there are appropriate enforcement mechanisms. Regular independent supervision, including from human rights organizations and civil society were mentioned as crucial. There could be an independent and neutral "**AI watchdog**", which could also provide an overview of AI development and collect best practices and bad examples of how not to deploy AI. Survey participants highlighted that an oversight body should have supervisory and quasi-judiciary competences. Participants emphasized that currently self-regulation is voluntary and cost-intensive, and might hence even have a competitive disadvantage. While the extent of oversight can depend on the service's size and impact, on the purpose of the AI deployment and its likely impact, auditing should not only be internal.

It was also stated that audits should cover both economic and tech aspects, such as AI's impact on human rights and measures taken to ensure user rights, to assess the quality of discourse, and AI's inclusiveness and remedies. In addition, oversight should be based on the **principles of neutrality, impartiality and human rights protection**. It was also highlighted that all stakeholders should be able to file complaints to an independent oversight, whose decisions should be binding. As models, audits for finances or sustainability were mentioned.

30. *What role could and should self-regulatory initiatives play? How can it be ensured that discussions around "ethical principles" are based on, and compatible with, human rights?*

Some respondents indicated that **self-regulation** should be the primary option as they are independent from government pressure and easier to react to new threats. At the same time, however, respondents emphasized that self-regulation alone is not sufficient and cannot replace **regulation and governance**. Industry self-regulation can only work under law and courts, participants stated, so that outcomes are verifiable and non-compliance draws legal consequences. Therefore, various respondents noted, AI principles should be enshrined in enforceable law. It was also noted that regulation often has unintended negative consequences.

Some respondents underlined that private discretion must be kept in appropriate limits and that self-regulation, adequate transparency and competitive neutrality should be complemented by multi-stakeholder engagement and regulation. International bodies could develop model guidelines for internal uses, and create indicators and principles for proper AI behaviour.

Survey participants also emphasized that self-regulatory initiatives should not facilitate unfair business practices or co-ordination and co-operation among services. There seemed to be

convergence that self-regulation must not be a self-made legislator or law enforcer. Self-regulation should be based on universally recognized legal norms, international human rights standards and ethical principles, but cannot substitute any of them.

Some respondents asked for a bigger role for civil society and international organization in self-regulation initiatives. Others called for co-regulation and other safeguards shared with civil society and academia.

It was also stated that while ethics alone are not sufficient and cannot replace legal safeguards, they can go beyond human rights, for example to ban manipulating peoples' behaviour through AI. Such ethical principles should be outcomes of multi-stakeholder consultations.

As a good example, **ombudspersons** for complaints, and self-regulation for the financial sector were mentioned, as was integrity and public trust.

31. *The #SAIFE Paper outlines how the use of AI-powered tools can create a chilling effect for the media, and can lead to self-censorship, particularly of marginalized voices, and to altered behavior in both online and offline spaces due to surveillance. How can AI-driven tools support the protection of journalists, and how can AI be beneficial for journalistic work and the media?*

Survey participants listed various AI-powered tools that can be beneficial for **journalism**, such as advanced tools to provide additional information and evidence, to research, or data journalism to detect biases or analyse huge amounts of data as used in tax haven reporting. AI could be used for translation and summarization, and for the editorial mission, e.g., by providing more diverse results and to better inform the audience. In this context, it was mentioned that public service media could develop AI and be open source, to better reflect editorial values.

Respondents highlighted that AI can raise the quality of research and investigation, but does not provide additional protection mechanisms. Other respondents suggested that AI could offer limited protection in conflict zones or for marking sources, safeguarding entities to automatically process certain procedures, or helping to register negative content such as co-ordinated smear campaigns.

32. *Do you have any other comments on the #SAIFE Paper and its preliminary recommendations, or would you like to raise additionally important aspects, areas of concern or sets of recommendations? Are there any other points you would like to raise?*

Respondents again referred to the power of a few intermediaries who can take down massive amounts of accounts and thereby, for example, block a full spectrum of political speech during sensitive times. Survey participants underlined the **need to treat different content differently**, as visual content, for example, is almost impossible for AI to properly assess, especially if it is of a sarcastic or artistic nature. Another issue mentioned was safeguards for live-streamed content.

Respondents also underlined the need to **find alternatives to monopoly tools** or data hungry solutions and the need to enhance **international co-operation** on similar initiatives. Respondents agreed that

there is no one-size-fits-all solution, which is why one must be careful not to over-generalize the challenges posed by AI.

33. *Do you want to add any specific observations in the context of the COVID-19 pandemic, and the tendency, as observed in the #SAIFE Paper, to revert to technocratic solutions, including AI-powered tools, which may lack adequate societal debate or democratic scrutiny?*

Respondents noted that the **use of AI increased during the COVID-19 pandemic**, for detection, review and potential removals. At the same time, some intermediaries suspended appeals, which means **limited redress** despite the increase in automated content moderation. Moreover, it was stated, the increase of AI-powered content moderation led to more mistakes in a time of increased dis- and misinformation. It was also highlighted that health is a sensitive field and speech might require more human involvement and review. Respondents underlined that the current level of transparency was not adequate, impacted users were not notified and there were no timely and robust appeals.

At the same time, respondents underlined that automated tools can help, e.g., by publishing rules of behaviour to decrease the level of panic. However, its use to combat the pandemic involves many questions, such as proximity tracing apps if there is no specific target or contingency plans despite high public funding.

Respondents highlighted that **emergency legislation** must be based in law and be necessary, proportionate to its purpose, non-discriminatory, temporary, focused, subject to regular review and the least intrusive approach possible – and that democratic oversight must be maintained in times of crisis. Survey participants also highlighted that it will be crucial to analyse the effects of emergency regulations and the increasing use of AI during the pandemic to propose better ways to protect freedom of expression in the future.