ANNEXURE B: Access to Information and Media Policy subcommittee report

How social media can support election integrity and media safety in 2024?

Answers arise from a "shadow" risk-assessment.

Executive summary:

As preparation for the South African elections, the SA National Editors Forum (Sanef) is leading an initiative with Media Monitoring Africa (MMA) to pinpoint risks for social media's role during the upcoming national polls. There are eight risks, grouped within four categories. They all have great relevance to the role that South African news media can play in the elections within the wider information ecosystem.

For each risk, both the impact and the likelihood are estimated, enabling a decision about which risks therefore should attract the most attention (for instance. the most impactful and the most likely merit the most attention). Going further, there are a number of actions that social media companies are suggested to take to "mitigate" the risks according to the tiers of priorities assessed. Lastly, there are "metrics" that set out measurable targets so that the effectiveness of these mitigations in practice can be meaningfully monitored.

All South Africans have an interest in helping ensure an optimum information environment for peaceful, credible and human rights-respecting elections, protecting the integrity of the elections and the online safety of journalists. Within this, what happens on social media is also of direct and immediate relevance to South African media.

Background:

On 23 January 2024, Sanef and MMA convened a workshop in Johannesburg, attended by representatives of the IEC, Press Council, Africa Check, media academia, lawyers, the Human Rights Commission, Digital Forensics Lab, journalists and editors from different publications. The task was to brainstorm about the risks - based on past experiences as well as foresight - as to the role of social media platforms in the upcoming national elections. This document is the fruit of the workshop discussions. Its mantra is "to be forewarned is to be forearmed".

After consideration by the Sanef council on February 10, the vision is that a final version be prepared to be published on the Sanef website, and brought to the attention of political parties, voters, the Independent Electoral Commission (IEC) and the general public. There will be an invitation to social media companies to then dialogue with Sanef and MMA. The aim thereof is to establish clearly what the South African public can expect of the platforms for the election period, and to compare this to the risks identified and outlined below.

An Appendix below spells out the specific motivation for this initiative and its foundation in international standards. The partners undertake to monitor the role of social media during

the election period, preferably in partnership with these companies and with access to data, in order to enhance positive actions and reduce potential negatives.

Sanef and MMA further pledge to take appropriate steps to put a spotlight on those actors who use social media services to endanger freedom of expression, access to information, safety of journalists and electoral integrity.

Overview:

Type of risk identified for the elections	Impact	Likelihood	Priority
To freedom of expression:			
● Silencing online public voices by intimidation	Medium	Medium	B
● Journalists attacked online	High	High	A
● Incitement via social media	High	High	A
To access to information			
● Disinformation on election	High	Low	C
● Manipulated media	High	Medium	B
To electoral integrity			
● Online attacks on electoral integrity	Medium	Medium	B
● Hacking and impersonation of IEC social media presence	High	Medium	B
Other risks			

In broad terms, Priority A should entail a platform having dedicated personnel with authority to act, and who can respond swiftly to the risk at hand. For Priority B, the platform should be proactively communicating with users and stakeholders, and also have a monitoring system in place. For Priority C, the platform should have plans in place - should these possible threats materialise.

In summary:

Priority	Type of risk	Broad preparation	Communicate results to public
A	1. Journalists attacked	Dedicate staff to act	√
A	2. Incitement via social media	Dedicate staff to act	√
B	3. Silencing voice by intimidation	Monitor	√
B	4. Hacking and impersonation of IEC social media presence	Monitor	√
B	5. Manipulated media	Monitor	√
B	6. Attacks on electoral integrity	Monitor	√
C	7. Disinformation on election	Back-up plan	√
	8. Other risks	Keep on radar	√

Risks, mitigations, metrics:

1. **Top-level risk: Journalists (and especially women journalists) are singled out for attacks (Priority A):**

Types of mitigations needed from platforms	Metrics
● Each platform sets up dedicated communications channels with journalists' associations and support groups for recognising attacks (eg. doxing, death threats, rape threat, brigading, gender-abuse), and for trusted flagging of attacks.	Email addresses and phone numbers of "focal points" are supplied to monitoring and reporting partners; An overview of this action is included in at least one transparency report before election day.
● Platform invites journalists (broadly conceived) who believe they may be at risk to volunteer their names (for example, on Facebook), or to informally provide names to trusted flaggers to give special monitoring and support	A confidential list of names of journalists likely to face attack is compiled, with input from MMA and Sanef; these accounts are then proactively monitored by dedicated staff on the platform, and actions taken accordingly to protect the journalists and prevent impunity for their attackers. An overview of this action, with granular data on the types of attacks (as per platform categorisation) and the corresponding actions taken, is included in at least one transparency report before election day.
● There are dedicated staff on the platform to monitor attacks on journalists through targeted sampling by crowdsourcing, and liaison with the MMA's MARS initiative.	Target: minimum of 20 cases detected and monitoring is kept up to date. An overview of this action (details can be confidential) to be included in at least one transparency report before election day.
● Capacity is put in place for rapid response with expedited redress when a journalist is attacked, ensuring protection for the victim, and ending impunity for attackers.	At least 50% of detected cases of attacks on journalists are effectively addressed, and information on "turn-around time" for such actions is provided in at least one transparency report before election day. Data to be disaggregated in terms of livestreams, 12 hour interventions, 24 hour interventions, and 48 hour interventions; data to be disaggregated in terms of attack types.
● Access to data is given to investigative journalists and credible researchers to analyse attack frequencies, themes, and networks.	Access should be given to at least 3 journalists and 3 research entities; Information giving the overview of this access should be included in at least one transparency report before election day.

2. **Top level risk: Expression is abused for hate speech, to foster polarisation, and to incite violence and/or scapegoating: (Priority A)**

Types of mitigations needed from platforms	Metrics
● Actively promote to users the platform's policies against hate speech and incitement to violence	3 campaigns with visibility to South African users, using 4 South African languages, and presented in easy-to-understand language.
● Reference the SA Equality Act for its definitions of hate speech sub-categories, and apply accordingly	Information on this, and training of staff in understanding and application of law, to be included in at least one transparency report before election day.
● Monitor (with partners) for racism, xenophobia and threats against legitimate exercises of the right to association, being sensitive to local context and languages	Dedicated team integrates key terms and phrases into dictionaries of smaller languages.
● Staff are dedicated to monitor intersections with offline dangers, such as illegal gatherings and co-ordination of violent physical attacks.	Information to be included in at least one transparency report about the platform's action taken to systematically look out for links between the platform's content and disruptive events as reported in media and statements by law enforcement.
● Assess for this risk in any inauthentic coordinated behaviour and expose such.	Information to be included in at least one transparency report before election day; more often if justified.
● Assess for algorithmic amplification and monetization and dial back algorithms and revenue sharing accordingly.	Information to be included in at least one transparency report before election day
● Co-operate with law enforcement, the Gender Commission, Human Rights Commission, Independent Electoral Commission, Electoral Court and the Equality Court.	This information to be included in at least one transparency report before election day
● To combat this risk in live streamed content, have dedicated capacity to predict, monitor and act in real time	Overview information on this mitigation to be included in at least one transparency report before election day

3. **Mid-level risk: Attacks aimed at intimidating people and silencing them from exercising free expression online. (Priority B)**

Types of mitigations needed from platforms	Metrics
● Proactively promote platform terms of service on relevant communications during the election period.	3 campaigns with visibility to SA users, using 4 South African languages, and presented in easy-to-understand language.
● Promote the SA Human Rights Commission's Social Media Charter, which guides users about their rights and obligations online.	All platforms meet with the HRC on this topic
● Stress-test relevant content policies to see if they, and their implementation, are fit for adequately distinguishing robust political debate from dangerous attacks that go beyond limits.	Two tests, one of which is commissioned from independent actors
● Monitor intimidatory attacks (with partners), and act against attackers.	At least 3 monitoring partners are identified and relationships developed; Disclosure of the percentage of total cases surfaced, in regard to which the platform then took further action (broken down into granular categories of actions – eg. warnings, downranking, sharing limits, deplatforming) Information to be included in at least one transparency report before election day
● Assess for inauthentic coordinated behaviour and bots in such attacks, and act against these, including exposing such	Results are included in at least one public report, issued before election day
● Assess for algorithmic amplification and (where applicable) monetization of attacks and dial back algorithms and (where applicable) revenue sharing accordingly.	Results are included in at least one public report, issued before election day
● Ensure functional communications-channels and staffing to receive and evaluate reports of attacks and take timeous action.	Email addresses and phone numbers of focal points are furnished to the partners who monitor and flag attacks; Results are included in at least one public report, with data on the "turn-around time" of acting on the different types of attacks. Data to be disaggregated in terms of livestreams, 12 hour interventions, 24 hour interventions, and 48 hour interventions.

4. **Mid-level risk: Hacking of IEC pages and other election related pages on social media and related communications channels (Priority B)**

Types of mitigations needed from platforms	Metrics
● Work with the IEC and political parties to address this in party codes of conduct and promote that they all use enhanced digital security.	Information to be included in at least one transparency report before election day.
● Have a plan for responding, and ensure capacity to respond to cases, where the pages/posts of IEC, Government and political parties are hijacked, or imposter activity is uncovered.	Information about incidence, and any results about actions taken, are included in at least one transparency report, issued before election day
● Proactively monitor use of IEC logo, and stop fake pages	Proactively remind users about the authentication process about which users' identities are verified, and integrate this into at least one visible media-and-information literacy campaign before the election.

5. **Mid-level risk: Disinformation and/or hate speech uses manipulated media such as synthetic content, cheap- or deep-fakes or AI generated content, and undisclosed sponsorship (Priority B):**

Types of mitigations needed from platforms	Metrics
● Pro-actively promote any content policies relevant to synthetic media (eg. policies requiring users/advertisers to disclose use of AI-generated content).	Include in at least one transparency report before election day, information about how the platform has advised users on relevant terms of service.
● Promote media and information literacy on this topic	One pedagogical exercise per platform, is promoted at least 4 times on the service
● Train fact-checkers in advance and have them on standby to check suspected fakes	At least 50 fact checkers per company, covering the 4 major languages, are trained, and overview information is provided in at least one transparency report before election day.
● Add labels, links to credible news and official sites to synthetic content when it is identified yet allowed to remain on the platform	At least 10 cases per platform, with the total number of cases, broken down into significant categories (eg. audio, video, imagery, text), with this information included at appropriate points during the election period, and in at least one transparency report before election day.
● Include this risk in the assessment of inauthentic coordinated behaviours	Any incidences to be reported periodically, and in at least one transparency report before election day
● Assess for such potentially harmful content being linked to algorithmic amplification and monetization.	Results are included in at least one transparency report, issued before election day
● To stop hidden advertising, pre-empt influencers not disclosing payment	2 warnings to 400 influencers (accounts with more than 7 000 followers) early on, about the

	relevant rules and disclosure requirements. Results are included in at least one transparency report, issued before election day

6. Third-level risk: Unjustified attacks on the credibility, fairness, competence and professionalism of IEC and/or its staff (Priority B)

Types of mitigations needed from platforms	Metrics
● Give prominence to IEC where it can make its own case and provide corrections and responses.	Include specific information on implementation of this mitigation in at least one transparency report before election day.
● Give prominence to credible news media, including by explaining fact-check corrections.	Include specific information on implementation of this mitigation in at least one transparency report before election day.
● Have a plan to monitor (with impartial partners) such attacks - including in comments on the social media pages/feeds of the IEC and political parties, as well as political advertising.	If this risk is materialised, activate plan and include information in at least one transparency report.
● Enlist fact-checking where facts are at stake, and promote labels and corrections - including circulating to those previously exposed (whether on social media, or on messaging channels).	Include specific information on implementation of this mitigation in at least one transparency report before election day.
● Assess for inauthentic coordinated behaviour and bots in amplifying this kind of attack, and expose such.	Timeously expose results of this mitigation , and in at least one transparency report before election day.
● Assess for algorithmic amplification and monetization and dial back accordingly.	Include specific information on implementation of this mitigation in at least one transparency report before election day.
● Have a position about potential collaboration with stakeholders (eg. human rights defenders, journalists, citizens, IEC staff, political parties, academics), who may wish to bring a case before the Electoral Court.	Implement as per position, and include results in at least one transparency report within the whole election period.
● Share cases, trends and patterns with the IEC	Metric: As needed, a minimum of 15 serious surfaced cases are addressed, with action taken on 10. This information to be included in a transparency report published during the election period.

7. **Third-level risk: Disinformation distorts access to authentic information and creates confusion about official electoral rules and arrangements (Priority C).**

Types of mitigations needed from platforms	Metrics
● If AI and partners flag that this risk is materialising, including in political advertising, there is a plan for responding, including assessing for cumulative disinformation narratives.	Activate plans to counter such disinformation through appropriate steps, and to uprank authoritative information. Publicise incidence and results as this may arise, and in at least one transparency report before election day.
● Ensure that systems monitoring for inauthentic coordinated behaviour include this particular risk.	Timeously expose these kinds of information operations that violate electoral integrity, and include in at least one transparency report before election day.
● Ensure that in the case of such inauthentic coordinated behaviour uncovered, there is also assessment for algorithmic amplification and monetization, and actions taken accordingly.	Such information should be in at least one transparency report before election day.
● Support and amplify genuine voter education campaigns	At least 2 visible campaigns during the election period, in 4 major South African languages
● Have capacity to investigate persistent falsehoods that jeopardise process integrity	Per company, there are plans to have a minimum of 50 trained fact-checkers on standby, covering at least the 4 major languages.

Other risks

Risks & mitigations needed from platforms	Metrics
For abuse on WhatsApp, the company should be monitoring through partnerships like with Real411	Receive and act on 300 complaints; Publicise overview in at least one transparency report before the elections. Data to be disaggregated in terms of turn-around times.
Group admins/convenors (on WhatsApp or Facebook) tolerate potentially harmful communications. The platform should empower Group Admins about the role they can play within company terms of service.	Metrics: circulate 2 messages to Admins about (a) their obligations, and (b) where they can find reliable information about the elections.
Where crisis situations emerge, platforms can add friction, reducing forwarding/sharing possibilities, and providing warning labels as applicable	Metrics: "Break glass" limits areimposed on forwarding/sharing to no more than 30 people at a time; this applies to a compiled list of persistent numbers where high-risk content is brought to light; Warning labels are standard on "borderline" content.
Persistent abusers violate terms of service without cease	Metric: Per platform, at least 50 cases identified and processed; 20 trolls deplatformed or face other sanctions depending on proportionality of restrictions
To avoid risk of content that promotes voter cynicism, platforms consider informing people about where to register and vote.	3 messages sent out on each platform about where people can find their polling station to register and vote.
Staged provocations that aim to attract restrictions in the interests of claiming victimhood, is a risk that can be mitigated by demonstrating bona fides in any restrictions imposed.	Metric: 2 reminders about platform ToS are sent out during the election period.

5. Conclusion:

Sanef desires to see reduced harms and increased benefits to the information ecosystem, as pertains to social media's role in the upcoming elections. Our members, their staffers, media sources and media audiences have a real interest in the information ecosystem of the upcoming elections.

Accordingly, the Forum along with MMA wishes to see:

- South Africa's hard-won freedom of expression and access to information being enabled on the platforms,
- Prevention of online voices of journalists and the public on these services being silenced, not least through cyber-misogyny.

Sanef further supports access to information, as distinct from access to lies and falsehoods. Sanef stands firmly against those who would use the election to scapegoat communities or to incite public violence.

The Forum has a vested interest in a functioning democracy, and related elections, as conditions for press freedom in particular and freedom of expression more broadly.

It is in this light that social media platforms are being urged to fulfil the civic obligations that come with doing business in a country. This means due diligence for the upcoming elections and taking account of media and civil society actors' proposals for advance risk assessments and mitigations. This kind of exercise can prepare both the companies and the public. It will serve to send notice to would-be disrupters and attackers that the major players within the communications infrastructure are in a high state of readiness.

Appendix:

- Sanef as a custodian of journalism in the South African information environment has a stake in the wider health of this environment and its impacts. Press freedom, the public's freedom of expression and access to information, and the safety of journalists, are core Sanef concerns. These issues are essential components for electoral integrity, and the health of the information environment plays a big role in this regard - with short and long term relevance to Sanef's mandate. Hence, while making its own contribution to the information environment, Sanef has a very direct interest in what social media platforms will contribute, including their plans to mitigate threats to the values we stand for.
- Human rights due diligence by private sector companies, which may include risk assessments are called for by the UN Guiding Principles on Business and Human Rights. Social media companies in the Global Network Initiative have expressed commitment to protecting a number of rights. UNESCO's guidelines for governing social media state: "Digital platforms should recognize their role in supporting democratic institutions by preserving electoral integrity. They should establish a specific risk assessment process for the integrity of the electoral cycle in the lead-up to and during major national election event". Such assessments are urged in the Principles and Guidelines for the use of Digital and Social Media in Elections of the African Association of Electoral Authorities.
- This background highlights the case for social media platforms to be doing evidence-based risk assessments ahead of the upcoming national and provincial elections, to be implementing mitigation measures for anticipated harms, and to be communicating on the impact.
- A summary "shadow" risk assessment for social media, as in this document, can help to encourage platforms to up their game in regard to the upcoming election.
- It is beyond doubt that online risks can translate into serious damage to the rights at the heart of elections and beyond, and Sanef will do its utmost to counter this outcome.