DeepFakery: Combatting illegal content, inaccurate information and fake news

Laurie Less and Tebogo Umanah

The evolution of social media poses both an opportunity and a threat to society creating vast learning opportunities and simultaneously gross human rights violations or what we call societal fault lines. The internet, mobile phones and other gadgets have increased the likelihood of users of social media, especially children, to come across illegal and harmful content. What exacerbates the problem is that information expressed in social media such as twitter is done so without any organizational or institutional filters, hence some of it can easily incite violence or cause harm, especially if it is propagandist in nature.

Fake new is defined as, "fabricated information that mimics news media in form but not in organisational content or intent.[1]" Fake content lacks the media's editorial norms and processes for ensuring accuracy and credibility of information and often escapes the regulatory teeth designed to protect citizens. It is illusory and misleading, amounts to misinformation and often times propagandist. More recently we experienced fake news abused during electioneering as a point scoring mechanism and in South Africa at the start of the Covid pandemic, Minister of Communications, Stella Ndabeni-Abrahams on 26 March 2020, gazetted requirements for the dissemination of COVID-19 information to citizens. This was as a response to the proliferation of fake news and inappropriate covid19 content.

During the month before the 2016 election in America, the average American encountered between one and three stories from known publishers of fake news. The challenge with false information on various social media platforms, is that it gets liked, retweeted and shared by thousands of people, spreading fake news and disinformation in a second. "Fake news is misleading and at its worst is an attempt to undermine national security", says Tambini, 2017.

So, to those of us who are still questioning whether they have come into contact with fakenews and or deepfakes: TIkTOK, DeepArt, Face Swap ring a bell? The reality of 'deepfakes' infiltrating every facet of your life has arrived.

How are deepfakes created?

Deepfakes is a hybrid of "fake" and "deep learning". Using AI artificial intelligence, deepfakes is a technique for **human image fusion**. The method combines and overlays existing images and videos clips onto source imageries, or videos, using a machine learning technique known as '**generative adversarial network'**. So picture this, if you have produced enough sound bites or videos and it's been distributed on various social media platforms, these technicians or perpetrators are able to cut, strip and combine pieces to create a complete new piece of content – featuring you in possibly very ominous situations.

More recently, since 2017, there is a proliferation of deepfake apps, a relatively new phenomenon in the realm of the internet age. Initially, deepfake technology was used at research and academic settings and by development amateurs. It then soon spread to the political sphere, aiming to alter peoples' perceptions of political leaders. This was soon followed by entering the entertainment realm. The use of deepfake

[1] Lazer et al (2018)

techniques was seen by millions of people around the globe in 2016 in Rogue One for the acting of Princess Leia, and in 2018 in Solo: A Star Wars Story, when Harrison Ford's face was inserted onto Han Solo's face.

The list of deepfake apps is endless and includes apps such as DeepFaceLab, Face Swap Live, Deep Art and AvengeThem and is available to anyone including a child, who has access to the internet. These apps have an ingenious capability to allow anyone to, within seconds, replace the original face in a video with someone else's face as well as to change your voice. **Deeptrace Labs**, a company that researches and detects deepfake apps, found that since 2018, the number of deepfake videos increased by an alarming 84%.

So, what's the fuss?

1. **Political interference and the erosion of democracy**

In the context of an election, fake news tends to undermine legitimate opposition[2]. It is a threat to democracy and the freedom of expression. It infringes upon the freedom to dignity of the people that it targets. Ironically, this harmful content attracts a lot of public attention because people are not immediately sensitised about the inaccuracy of the information.

Fakenews has the ability to erode and undermine the very fabric of our various global institutions, and country to country associations. What should be considered is the development of a single regulatory framework, a SADC, AU and or BRICS convention with associated resources as a go to – when assistance is required. Because the critique at hand is that FAKENEWS play a critical role or is a form of undermining or interfering in the sovereignty of our member countries. We should be concerned as members of a global society of the possibility of undermining the sovereignty of states and therefore the voice of these countries **MUST** be vigilant to form a united harmonized defense that will discourage the growth of these tendencies. How will we hold each other accountable to this harmonized regulatory framework or a convention? We need to appraise and evaluate the damaging effect of fakenews on national state sovereignty. The use of fakenews is subversive and contrary to the essence of the nation state. The establishment of global economic blocs such as BRICS is in and of itself a desire for autonomy (self-determination) and not being beholden to hegemonic powers.

Imagine how easily fakenews could serve inappropriate persons to sway the views of a nation; the potential for damage is endless. We will be remiss in our responsibility as nation states to defend our citizens and to encourage gains made in building a global citizenship - where people can move as efficiently as goods across borders and international solidarity becomes a reality.

It is therefore critical that as member countries we monitor fakenews and the manner in which we can prevent it legislatively as well as PUNISH it.

2. **Societal fault lines:**
    a. **Fakenews used to stir up emotional fault lines that lead to violence and genocide.**

Twenty-six years ago (April), all hell broke loose in Rwanda, hordes of members of the Hutu ethnic majority, armed with machetes, spears, nail-studded clubs, and other rudimentary weapons, moved house to house in villages, hunting for Tutsis, the second largest of Rwanda's three ethnic groups. The radio station RTLM, together with leaders of the government, had been inciting Hutus against the Tutsi minority, repeatedly describing the latter as inyenzi, or "cockroaches," and as inzoka, or "snakes." The

[2] Tambini, 2017.

radio station regrettably had many listeners. We suffered the results of ethnic cleansing and genocide. The potential for a radio programme to wipe out a nation became real.

b. Call of Duty game Black Ops recklessly promotes conspiracy theories:

Just recently in The New Yorker, developers of the game, "Call of duty: BLACK OPS COLD WAR", abused a video clip of a Soviet Union defector without contextualizing it (in the games trailer). It used a real interview by Russian defector Yuri Bezmenoz, that occurred 1984 (during the cold war) - to promote a new addition to the game BLACK OPS COLD WAR. Yuri is shown stating how the USA will gradually be undermined to allow for minorities such as African Americans, women and the LGBTQI communities to destabilise American society.

Yuri warms Americans about the move of these social movements – he infers that the Soviet Union has a role in destabilising American society. Many Americans who have seen this or read this have taken this as gospel truth. So, think about this: an impressionable 18-year old with access to a gun listens to this, plays the game Call of Duty Balc Ops and potentially acts on this…. Case in point the Norwegian massacre, where a young man watching Call of Duty Black ops II (2012) went on to massacre 77 people.

This is how popular culture is used to insidiously normalize abnormal behaviour such as providing political rights to various groups. This is how unprogressive ideas are sold to the public and more ominously where gaming and internet platforms are used to stir up hatred or identity fraud.

3. The internet has no firewall for patriarchy[3]: tools that become weaponized against women

When powerful women stir the hornets' nest they are most vulnerable to vicious online attacks. Women bear the brunt of these social outcries. The capability to reinvent a character allows for the creation of endless hoaxes, fake news, nude and pornographic scenes and revenge porn, and target vulnerable individuals such as girls and women. Mutale Nkonde, a fellow at the Data and Society Research Institute in New York, states that "The DeepNude App (a deepfake app) proves our worst fears about the unique way audio-visual tools can be weaponised against women", ultimately altering our perceptions of people and controlling women's bodies. Many celebrities, politicians and alarmingly non-celebrities have experienced this, as deepfakes software can be purchased online for as little as R45.

Children and other citizens globally are exposed to misinformation and other forms of illegal content, such as child pornography. The capability to reinvent a character allows for the creation of hoaxes, fake news, nude and pornographic scenes and **revenge pornography**. The use of Deepfake apps tends to perpetuate cyber-misogyny, as women remain vulnerable to manipulation of their pictures by those acting maliciously in cyber space. Social media is a key conduit of fake news; and given the access to cellphones and other gadgets in the 4th Industrial Revolution era, it becomes crucial to regulate such content as it remains harmful to vulnerable groups, mainly women and children.

Initially, deepfakes appear to be innocent and exciting, but the individual could be lured into the destructive side of the app. The result could be an individual who is ill-informed about the harmful effects of their actions, not only on themselves, but to those that they target through using deepfake apps.

More disturbingly, Deepfake Apps target especially our children. Apps form a key element of our children's current social lives and is evident in their beliefs, their value systems and their behaviours. This all occurs

[3] Term coined by Mr Oupa Makhalemele, Senior Researcher at the FPB.

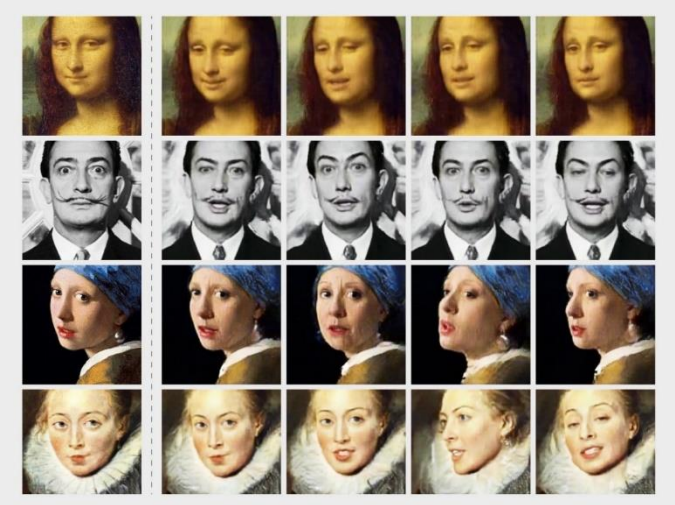
while children are physically, and also emotionally and psychologically in their developmental phases. In the end the use of Deepfake Apps amongst children may result in their inability to differentiate between reality and fake information; the ruining of relationships with significant others, their reputations as well as their online reality with cyberbullying forming a key part of this constant element in their lives.

The manipulation of images and videos using artificial intelligence has the ability to become a destructive mass phenomenon. Arwa Mahdawi of The Guardian believes that the key factor behind the creation of deepfake pornography is the desire to humiliate and control women. An artificial intelligence researcher, Alex Champandard, opines that due to the inability to differentiate between reality and fake media, humanity has entered an age in which it is impossible to know when content represents truth.

Taking Action against the abuse of information and fakenews

Globally numerous governments, such as the United Kingdom, EU, USA as well as SA close down deepfake apps and make the creation and distribution of deepfake Apps punishable by law.

Locally in SA, the Film and Publication Board (FPB), through the implementation of its amendment Act, 2019, the 2017 Cybercrimes and Cyberbullying Bill, aim to rationalise the laws of South Africa which deal with cybercrime and cybersecurity and to criminalise the manufacturing and distribution of malicious communications as a means to provide interim protection measures.



We acknowledge that there is an increased demand for online content and technologically advanced content creation, therefore the concomitant demand for content regulators in BRICS to increase its monitoring of digital platforms and social media. Arguably deepfake apps will affect the way we perceive life and place further pressure on our societal norms and values. But the key concern should be the effect deepfake apps have on women, vulnerable minority groups and children.

(Source: https://www.wired.com/story/deepfakes-getting-better-theyre-easy-spot/)

Real 411 is a prime example of a civil society led initiative in South Africa, aimed at addressing harmful online content. It emerged out of a need to curb fake news by setting up a system that would enable members of the public to report misinformation. Real 411 has a code of conduct, which applies to offences that comprises harmful false information, hate speech, incitement of violence and harassment of journalists. The code of conduct established the Digital Complaints Committee (DCC). The code of conduct seeks to strike a balance between competing rights and interests by taking into account factors such as freedom of expression, satire and public interest. Additional recourse for the public is to apply to the equality court for relief, approach the South African Human Rights Commission for assistance and also, to check a fact-checking organization for verification.

Legislative Frameworks in place to combat distribution of illegal content including fake news

South Africa has a few regulatory frameworks that seek to combat the distribution of illegal content and inaccurate information, including fake news.

- The Constitution of the Republic of South Africa (1996);
- the Film and Publication Act no 65 of 1996 as amended. The Amendment Act of 2019 was amended in order to criminalise distribution of illegal content, including distribution of inaccurate or fake news as well.
- The Disaster Management Act Regulations makes provision for prosecution of anyone who creates or spreads fake news.
- The Preventing and Combating of Hate Speech Bill of 2018.

Hate speech has become dangerous in the age of social media, where people freely express their opinions in the public arena.

Means of regulation

There are three possible mechanisms for regulating distribution of illegal content which could be considered. There is either self-regulation by the platforms, government intervention or co-regulation.

Direct government regulation

This phenomenon entails the establishment of rules by the state to regulate private business. Direct government regulation, however, could be construed as censorship, given constitutional rights of citizens in different nation states. In South Africa for example, the bill of rights protects the right to freedom of speech and expression, albeit with certain limitations. However, government regulators may not easily maintain objectivity in defining and imposing some regulations. They may end up developing regulations that suit their political interests.

Governments by their nature have a responsibility to balance public interest by ensuring a free flow of news and simultaneously protecting the dignity of their citizens through its legislative regimes. In South Africa, direct lawsuits have taken place, where people who felt that their characters were defamed through social media approached the equality court to seek justice. Therefore, law on defamation becomes crucial as a means to control fake news that is defamatory and harmful.

From Self-Regulation to Co-regulation

Co-regulation refers to a combination of self-regulation and government regulation, where consensus is reached between actors in the regulatory space such as judges, legislators, civil society and the regulatory authority The Film and Publication board in this case. This process enhances the legitimacy and efficacy of regulation since interests of all actors are considered[4]. The FPB uses a co-regulatory model working with industry, government and the public. The laws and policies are in place for public accountability. Whilst industry ensures the application and compliance with the law in the instance of a transgression an independent enforcement committee will review the transgression.

Another option is **self-regulation**, which allows industry and in this case, the media, to have regulatory mechanisms in place such as code of conduct, rules and standards that an industry should comply with as well as mechanisms to monitor compliance to the set rules and standards. However, self-regulation of

[4] Lievens and Dumortier, 2005.

the internet has been criticized due to problems of legitimacy and accountability, lack of credibility and transparency and concerns about protection of freedom of expression. Self-regulation has also been criticized for the lack of enforcement of sanctions.

The Press Council, the Press Ombud and the Appeals Panel are an independent co-regulatory mechanism set up by the print and online media to provide impartial, expeditious and cost-effective adjudication to settle disputes between newspapers, magazines and online publications, on the one hand, and members of the public, on the other, over editorial content of publications. It is based on two pillars: a commitment to freedom of expression, including freedom of the media, and to high standards in journalistic ethics and practice. The South African Press Code guided journalists in their daily practice of gathering and distributing news and opinion and to guide the Press Ombud and the Appeals Panel to reach decisions on complaints from the public.

Conclusion

In conclusion, our original objective as the AU, SADC and BRICS was to improve our global economic position, reform our financial institutions, through building an integrated, prosperous and peaceful Africa, driven by its own citizens representative of a dynamic force in the global arena[5].

We should consider moving towards harmonization of a single regulatory regime or a **CONVENTION**, where a **single code of conduct** may be adopted across member countries. This may call for a network of global or multiple players brought together to tackle these challenges and participate in a single harmonized body to regulate media content. Principles of accountability, transparency and respect for human dignity, and the safety of the Child - amongst others, may be considered and adopted as guiding principles for such a convention and or regulatory framework. The recent development of a single BLOC in Europe to combat CSAM is , The European Commission adopted a proposal for a Regulation on a temporary derogation from a number of provisions of the e-Privacy Directive as regards the use of technologies by number-independent interpersonal communications providers for the processing of personal data and other data for the purpose of combatting child sexual abuse online (effective December 2020). This could reduce many obstacles some technology companies faced in regard to scanning their services for hashes[6] of known CSAM. This is the result of the combined intensive national and international lobbying efforts of Hotlines, INHOPE, NGO's, industry and law enforcement over the past year. This demonstrates that it is possible when member states converge and agree on fundamental principles to combat fakenews – no matter how wide the geopolitical boundaries.

Our global agenda has been firmly rooted, but the overall wellness of a nation is measured in far more complex terms than its mere economy, when the essence of our democracies, societies and social interactions is being threatened and DEEPFAKES are an anathema to TRUTH…

then good governance, self-determination and autonomy will be measured in the currency of TRUTH into the future.

[5] AU mission statement.
[6] hashtags