



# Together We Talk

**Guiding Respectful  
Digital Dialogue**

A toolkit to learn and practice to counter dangerous speech and foster peace in digital spaces



MADE IN SUDAN  
FUELED BY SCIENCE

### Acknowledgements

We would like to express our profound gratitude to the UN Peacebuilding Fund for their invaluable support in developing this 'Together We Talk' toolkit.

This toolkit was prepared by Laura de Reynal, Alexandra de Filippo, Basma Gubara, Ali Muntasir, Amal Tahal and Jennifer Colville. We thank Josh Martin & Nour Nasr for their advice during this project.

Graphic design by Natalie Jane Worth.

This toolkit brings together resources, research, and practical activities inspired by the work of leading scholars from various fields. We deeply appreciate and acknowledge their significant contributions, and each has been cited within the toolkit. Our goal is to make their important findings accessible to a broader audience, providing practical tools that can help build a peaceful future. We're standing on the shoulders of these intellectual giants, aiming to popularize their work and inspire everyone to take action.

The views expressed in this publication are those of the author(s) and do not necessarily represent those of the United Nations, including UNDP, or the UN Member States.

UNDP is the leading United Nations organization fighting to end the injustice of poverty, inequality, and climate change. Working with our broad network of experts and partners in 170 countries, we help nations to build integrated, lasting solutions for people and planet.

Learn more at [undp.org](https://undp.org) or follow at @UNDP.

Copyright ©UNDP 2023. All rights reserved.

One United Nations Plaza, NEW YORK, NY10017, USA



# Together We Talk

## What is in this toolkit?

Upon completing this toolkit, with an active participation in the exercises and usage of the resources provided, you should be equipped to:

- 1 Identify various forms of Dangerous Speech, including misinformation, disinformation, and hate speech, prevalent in your communities.
- 2 Practice critical thinking before deciding to share content, using checklists and prompts.
- 3 Comprehend the cognitive biases and fallacies that make us susceptible to being both victims and unintentional propagators of Dangerous Speech.
- 4 Safely respond to Dangerous Speech using techniques such as pre-bunking, debunking, and counterspeaking.
- 5 Encourage others to join you in combating Dangerous Speech.

# Your learning journey

Learn how our minds work. Learn the power of a growth mindset and reflect on our values.

Learn to recognize and identify different kind of dangerous speech.

Learn to recognize and identify different forms of hate speech to prevent them from spreading further.

Learn to recognize different forms of mis-dis-mal information and understand why they spread.

Learn about the role of social media in the spread and the fight against dangerous speech.

Learn how to respond to dangerous speech in a scientific and proven way.

Learn how to set a S.M.A.R.T goal for your work.

Learn how to be safe online.

# Welcome to Together We Talk

So, what is dangerous speech? Why are we here?

You've likely come across it at some point, maybe even without realizing it. It stirs up controversy, instigates conflict, and yet can seem so innocuous wrapped up in a catchy meme or a trending hashtag. It's disconcerting, but you're intrigued.

**Exactly what makes speech 'dangerous' and can't we just scroll past it? Let's dive deeper.**

Welcome to "Together We Talk", a toolkit designed to navigate these murky waters, championing dialogue, unity, and peace amidst the digital tumult.

If you are a curious learner, a believer in peace and non-violence, an influencer, a leader, an organization dedicated to making the world a better place, or simply, a human, then this toolkit is for you!

## What will I need?

### MATERIALS

Scratch paper, pen/pencil, a quiet place.

### TIME

At your own pace. We recommend 60-90 minutes.



# Understand the power of your mind

Growth mindset: the power of "yet".

In this activity, we'll leverage the 'power of yet,' a key concept in cultivating a growth mindset, to work towards improving our ability to counter Dangerous Speech and educate others.

1

## Reflect on your journey



Start by thinking about your current abilities to identify and counter Dangerous Speech.



Write down one or two areas where you feel you're not as effective as you'd like to be.

### EXAMPLES:

*"I struggle to remain patient when encountering hate speech online."*

*"I have 0 empathy for perpetrators of dangerous speech"*

*"I tend to share content quickly and I don't know how to fact check."*

### YOUR TURN:

1. \_\_\_\_\_

2. \_\_\_\_\_

2

### Apply the 'power of yet'



Now, take these statements and add the word 'yet' at the end of each.

#### EXAMPLES:

*"I struggle to remain patient when encountering hate speech online, **yet.**"*

*"I have 0 empathy for perpetrators of dangerous speech, **yet.**"*

*"I tend to share content quickly and I don't know how to fact check, **yet.**"*

#### YOUR TURN:

1. \_\_\_\_\_

2. \_\_\_\_\_

3

### Set your goals



By adding 'yet,' we acknowledge that we are in the process of learning and improving, and that it's alright not to have mastered everything. Now, for each 'yet' statement, set a specific, achievable goal that will move you closer to mastering that area.

#### EXAMPLES:

*"I will use the checklist provided to think critically before I share the content."*

#### YOUR TURN:

1. \_\_\_\_\_

2. \_\_\_\_\_

Remember, change doesn't happen overnight, but by continually applying the 'power of yet,' you can cultivate a growth mindset and improve your ability to combat Dangerous Speech.



# We think fast, we think slow

“Thinking Fast and Slow” is a psychological framework developed by Nobel laureate Daniel Kahneman. It describes two ways our brain processes information.



Source: Daniel Kahneman

**1**  
The **“Fast” (System 1)** is our autopilot mode. It is instinctive, emotional, and often based on mental shortcuts. It’s the mode we are in when we react quickly, often getting distracted or led by our emotions.

**2**  
The **“Slow” (System 2)** is the careful pilot. It’s logical, analytical, and takes its time. It is the thinking mode we activate when we analyze a situation and make thoughtful decisions.

Click [here](#) to watch a short explanatory video & learn more.

### Did you know ?

On an average day, we're faced with approximately 35,000 decisions. That's a lot of information to manage! To manage our limited cognitive resources, we use numerous mental shortcuts, called heuristics. These help us navigate through the world, but they can also result in judgment errors.

### EXAMPLE:

Imagine If you're running a race and you pass the person in 2nd place, what place are you in?

System 1, our fast thinking, might immediately respond with "first place". However, if we engage System 2, our slower, more thoughtful process, we realize that by passing the person in 2nd place, we now occupy the 2nd place, not the 1st.

**So, why should you care about these when it comes to combating dangerous speech? What's the link to peace, unity or respectful online conversations?**

Here's the situation: when System 1 takes the wheel, it is super easy to impulsively share harmful content that fuels tension, hatred, and violence. It is like pouring gasoline on a fire.

Our goal is to empower you to take control, use your System 2 more often, and think more analytically and rationally. This way, we can promote thoughtful and respectful conversations online, even in the face of conflicting viewpoints.



### REFLECT

1. Think about a time when your 'Fast' thinking led to a knee-jerk reaction online. What was the situation and how did you respond? In retrospect, do you think 'Slow' thinking might have led to a different outcome?
2. Consider a moment when your 'Slow' thinking mode was on full display. How did taking a step back and analyzing the situation help you navigate a tricky conversation or decision? How did it feel to engage that analytical, rational part of your brain?
3. Consider sharing your insights and reflections with a friend to learn from them and help them learn from you too.

# Values and intentions

Grab a piece of paper, your phone notes app, or just start a fresh document on your laptop. If none of these are accessible, no worries – just find your thinking space.

There are no right or wrong answers in these exercises, this is about getting to the core of your values and goals.

### Did you know ?

Self-reflection boosts our self-awareness, helping us understand how our emotions affect our (online) behaviors.<sup>1</sup>

Identifying our personal values and goals can help us guide our (online) actions towards more respect and understanding.<sup>2</sup>

## Top Five Values Exercise



Write down your top five personal values. What truly matters to you? What principles guide your actions and decisions? These can be things like friendship, sports, honesty, family, creativity, justice, creativity or anything that truly resonates with you.

### EXAMPLE:

*"My top 5 values are: Empathy, Physical Health, Community, and Growth."*

### YOUR TURN:

1. \_\_\_\_\_
2. \_\_\_\_\_
3. \_\_\_\_\_
4. \_\_\_\_\_
5. \_\_\_\_\_

<sup>1</sup> (Sutton, 2016)

<sup>2</sup> (Bardi et al., 2009)

## The Five Whys Exercise



**EXAMPLE:**

"My intentions for opening this toolkit are to learn more about dangerous speech and how I can help counter it in my community."

Here's where we dive deep. The Five Whys exercise, developed by Taiichi Ohno at Toyota in the 50s, is a powerful tool for finding the root cause of something. It's an exercise that promotes critical thinking, prevents the recurrence of problems, and helps people develop problem-solving skills by asking deep questions.



Take the reason you wrote down for opening this toolkit and ask yourself 'Why?' five times.

*For example, if your initial reason was 'I want to understand dangerous speech better', your first 'Why?' might lead to 'Because I want to help my community'. Keep going until you've asked 'Why?' five times. This will help uncover your core motivation and shed light on the root cause of your intentions.*

**EXAMPLE:**

**Why** did I open this toolkit? *Because I want to learn how to counter dangerous speech.*

**Why** do I want to learn this? *Because I've seen how dangerous speech can divide communities.*

**Why** does that concern me? *Because I value unity and peace.*

**Why** do I value unity and peace? *Because I believe everyone has the right to feel safe and respected.*

**Why?** *Because I have seen the suffering on both sides, and I want it to stop.*

**YOUR TURN:**

1. \_\_\_\_\_
2. \_\_\_\_\_
3. \_\_\_\_\_
4. \_\_\_\_\_
5. \_\_\_\_\_

1

# Chapter 1

Learn about the different kinds of dangerous speech and how they spread

## Did you know ?

Being well-informed equips you to respond or 'counter speak' effectively, enabling you to spot early signs of dangerous speech which might be overlooked by others.

## Key message:

The internet is full of harmful narratives, which no matter their intent or perceived harm, share a common characteristic - their potential to incite violence and endanger communities. This type of rhetoric, which can be called "Dangerous Speech", applies to all forms of communication capable of sparking violent actions.

In this chapter, we will dive into various forms of dangerous speech, from misinformation and disinformation to hate speech. You will learn to analyze the way they are written and the reasons why they get shared and spread so easily online. We will equip you with strategies to critically analyze information before sharing it online.

**The details are crucial.** Being well-informed equips you to respond or 'counter speak' effectively, enabling you to spot early signs of dangerous speech which might be overlooked by others. This is not only about reaction, but also prevention. You, along with your community, play an instrumental role in preventing the spread and escalation of violence. **So, let's learn!**

BE THE  
CHANGE!

**Did you know ?**

Fear of 'the other'<sup>1</sup> is a complex issue that has been studied from various perspectives, including neuroscience, social psychology, and evolutionary psychology. There is early evidence to suggest that fear is actually an automatic or programmed reaction to the others, especially in periods instability and scarcity.

In addition to these evolutionary explanations, there is also evidence that fear of the other can be learned by associating negative information or experiences to their group.

<sup>1</sup> - Also called 'Out-Group' in Sociology: the people who do not belong to a particular in-group in a society.

# 1. What is Dangerous Speech ?



**'Dangerous speech'** is talk that portrays others as threats, justifying violence<sup>1</sup>. It can be any form of expression (e.g. speech, text, or images) that can increase the risk that its audience will approve or participate in violence against members of another group.



**No one is born with hate or fear**—these are taught over time. Across cultures and history, leaders have used narratives to demonize 'the other', and put groups against each other. The vocabulary might change, but the themes are very similar. We can recognize the patterns.



**Dangerous speech promotes fear and escalates risks of violence rather than directly causing it.** It is impossible to say that it directly causes violence, due to the subtleties of human behaviors, but these narratives influence people and therefore it is essential to understand them better and learn to manage their risks. As such, monitoring dangerous speech serves as an early warning system for potential violence.

**This approach is not about imposing censorship, but about educating individuals to be less susceptible to the influence of harmful speech.**

<sup>1</sup>- Benesch, Susan & Glavinic, Tonei & Manion, Sean & Buerger, Catherine. (2018). Dangerous Speech: A Practical Guide.

### What is the difference between dangerous speech, hate speech & misinformation?

There are different kinds of dangerous speech. You may have heard of some of them or come across them online. There are subtle but important differences between them. Here are some practical and useful definitions, refer to them when you get lost:

#### Did you know ?

Dangerous speech can present a risk to increase violence, regardless of the intention of the people sharing it.

|                         |                                                                                                                                                                                                                                                                                                                                                                    |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Misinformation</b>   | Refers to those who spread false information without realizing it, usually because their friends or others do                                                                                                                                                                                                                                                      |
| <b>Disinformation</b>   | False information deliberately and often covertly spread (as by the planting of rumors) in order to influence public opinion or obscure the truth                                                                                                                                                                                                                  |
| <b>Mal-information</b>  | Genuine information that is shared to cause harm. This includes private or revealing information that is spread to harm a person or reputation                                                                                                                                                                                                                     |
| <b>Hate speech</b>      | Language that generates division and hatred against communities based on their identity. Hate speech refers to the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender (i.e., sexism), their ethnic group or race (i.e., racism) or their beliefs and religion |
| <b>Dangerous speech</b> | Communication that may help catalyze mass violence by moving an audience to condone, or even take part in, such violence. It includes hate speech, dis/mis/mal information                                                                                                                                                                                         |

The key difference in these messages is the intention of the person sharing a message. Sometimes, people spread misinformation or hate speech unintentionally, being caught up in emotional narratives or sharing shocking content without verifying it. People may not be aware of the implications of their actions, or how such messages can fuel hostility and misunderstanding. Regardless of intent, the impact can still be harmful. This inadvertent spread of misinformation or hate speech is a significant concern in the fight against dangerous speech.



**Did you know ?**

False stories are more likely to be shared: research shows that false news stories are 70% more likely to be retweeted than true stories.

True stories travel slower: It can also take true stories about six times longer than false stories to reach people.<sup>1</sup>

1 - Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science

Exploring the 7 rhetorics with a critical thinking mindset

## 2. What is dis-mis information?

### How does it work?

There are seven types of mis- and disinformation messages<sup>1</sup> which can contribute to the spread of problematic narratives, potentially leading to confusion, polarization, and distrust.

1. **Satire or parody:** harm is not intended, but it has potential to fool and mislead people.
2. **Misleading content:** misleading use of information to frame an issue or individual.
3. **Imposter content:** impersonation of genuine sources.
4. **Fabricated content:** false content, designed to harm and deceive.
5. **False connection:** when headlines, visuals or captions don't support the content.
6. **False context:** genuine content shared with false contextual information.
7. **Manipulated content:** genuine information or imagery is manipulated to deceive.

### 1. Satire or parody

Satire or parody are forms of humor that often use exaggeration, irony, or ridicule to expose and criticize people's opinions, stupidity or vices, particularly in the context of contemporary politics. The problem is when they become (too often) mistaken as real information. Several satirical media websites or pages immitate real media websites, often creating confusion among readers. Even when the humorist adds a disclaimer to say that their story is a joke, some readers still believe and share the content.

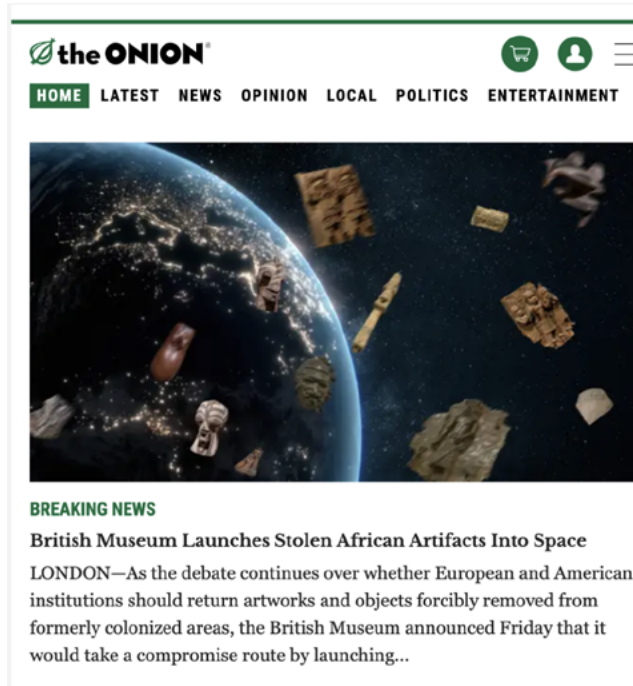
**EXAMPLE**

World Leaders Confess: All International Conflicts Actually Settled By Fortnite Matches. Winner gets to decide policy, loser has to dance 'Take the L'. #BattleRoyaleDiplomacy

This headline is obviously exaggerated and not meant to be taken seriously. However, some people may not be able to recognize this and might share the post without realizing that it's meant to be satire.

1 - Wardle and Derakhshan (2017)

### EXAMPLE



In this example from the Onion, the title is humorous and satirical. It is meant as a joke.

However, the topic is very real and political, as you can read in the first line "As the debate continues over whether European and American institutions should return artworks and objects forcibly removed from formally colonized areas..."

We can see how this kind of parodic content could mislead some un-prepared readers.

[Wikipedia maintains a list of satirical websites](#) that you can check out for more information on the topic.



### Learn to think critically

Use the questions/checklist below to practice critical thinking and become an expert at recognizing this kind of content:

- Does the headline seem too outrageous or exaggerated to be true?
- Is the website's "About us" section mentioning their goal to share humor?
- Is the tone of the post humorous or sarcastic?
- Are there any indicators in the post that suggest it is meant to be humorous or exaggerated, such as exaggerated quotes or over-the-top images?

## 2. Misleading content

Misleading content is information that can be either false or partially true, and presented in a way that gives a false impression or leads to incorrect conclusions. Misleading content can take many forms, including manipulated images or videos, false or misleading headlines, and biased or incomplete reporting.

### EXAMPLES



#### NEWS ALERT

**News Alert: Looting incidents skyrocket, all conducted by Purple People. Stay safe out there!**

A news article frames looting incidents as if they are exclusively conducted by a specific group, misleading readers to stereotype that group.



#### BREAKTHROUGH

**70 new COVID cases caused by schools reopening in France.**

A news article makes the misleading conclusion that 70 new covid cases appeared due to the school's reopening, while in fact, the students were "likely" to be infected before the schools opened. The content is not false: there are 70 covid cases, but the causal relation with the school opening is misleading.

**These headlines are misleading and false, but some people may take it at face value and share it without fact-checking.**



### Learn to think critically

- Is the headline sensational or attention-grabbing, but not supported by the actual content of the article or post?
- Is the source of the information biased or known to spread false information?
- Does the information presented seem too good or shocking to be true?
- Are there alternative sources that contradict the information presented in the content?
- Is the information presented in a way that is meant to manipulate or deceive the reader?
- Faulty logic: is the logical argument strong? Is it backed by science?

---

---

---

---

---

### 3. Imposter content

Imposter content is a way of sharing mis-disinformation that comes in the form of fabricated images and text, that imitates true media outlets, or organizations or people. This involves the impersonation of good and genuine sources and often copies their branding.

#### EXAMPLES



**Official Twitter Account**  
@Official\_Twitter\_Account

We are proud to partner with Bitcoin to offer free investments to the first 500 people who sign up ! Enter your email here to participate.

A fake account pretending to be Twitter's official account tricks people into sharing private information in exchange for free bitcoins.

<http://www.education.gouv.france.fr>

Ensure your children get access to a free iPad for school. We will send penalties to the ones who don't sign up for the program. [Sign up here](#).



A website that has a similar URL than the official "gouv" website for the government tricks parents into sharing personal information about their families.



**Learn to think critically**

- Is the source of the content different from what you would normally expect?
- Are there any signs that the content has been altered or manipulated in some way?
- Does the content contain any unusual language or formatting that is atypical for the source?
- Is the content being shared through an unusual or unexpected channel, such as an email from a friend's account that seems suspicious? Imposters often use usernames that are similar to that of the official account.
- Are there any red flags that suggest the content may not be legitimate, such as a request for personal information or a demand for money?

---

---

---

---

---

---

---

---

---

---

### Did you know ?

Deep Fakes are videos of people that have been modified with the latest technology to appear to say something that they are not saying. They are typically used maliciously or to spread false information.

Deep fakes are getting harder to recognize, due to the advance in modern technology.

For example, see this [very realistic deep fake of Morgan Freeman](#), a Hollywood actor. Watch until the end to see the actor who is actually the real person speaking.

[In this other interesting example](#), we see Mothers saying toxic things to their daughters about beauty. Deep Fake technology has been used to make beauty influencers look like the mothers, and it's very realistic!

## 4. Fabricated content

Fabricated content is completely false, designed to deceive and cause harm.

### EXAMPLES

### Pope Francis endorses Donald Trump for USA Presidential Elections

This is an example of a completely fabricated story that was widely shared and believed by thousands of people.

#### Example of a fabricated video, with the same footage and a different audio:

1. [Original video: link](#). In this original video, we see humanitarian aid refused and left on the floor, because the recipients refused to be photographed to receive the aid. They found that request to be humiliating and did not want to be a part of it.
2. [Fabricated or modified video: link](#). In this modified video, the images are very similar, but the audio has been fabricated and taken from another video. We hear citizens thanking the king Salman of Saudi for the generous donation. The sound is from 2020, while the footage is from 2023.



### Learn to think critically

- Are there any red flags that suggest the content may not be legitimate, such as an unusual or sensational headline?
- Is the information presented supported by reliable sources?
- Are there any attempts to deceive or manipulate the reader, such as by presenting fabricated evidence or false statistics?
- Is there any conflicting information from other sources that contradicts the claims being made in the content?
- Does the content use language that is designed to incite hatred or violence against a particular group of people?
- Does the content come from a website known for its reliable information?

**5. False connection**

False connection content is when headlines, visuals, or captions do not support the content of an article or story. A common example of this type of content is clickbait headlines, where the headline is designed to attract attention and encourage clicks, but does not accurately reflect the content of the article. False connection can be particularly insidious, as it can use misleading visuals or headlines to suggest a connection between two things that are actually unrelated.

**EXAMPLES**

A social media post shows a photo of a group of people protesting in the streets of Sudan, with a caption that claims they are protesting against a particular ethnic group. However, upon closer examination, it becomes clear that the photo is actually from a completely unrelated protest in a different country.



Look at this picture—Sudan is a war zone. We need to stand against this violence.

An article about violence in Sudan uses a photograph of violence from another conflict, falsely suggesting the situation in Sudan is worse than it is.



This politician was spotted at the violent incidents yesterday. They are clearly involved!

A post associates a politician with violent events they had no involvement in, creating a false connection and smearing their reputation.



### Learn to think critically

- Does the headline or visual presented actually match the content being presented?
- Are there any red flags that suggest the content may be misleading or manipulated?
- Is the context of the information presented clearly explained, or is it left up to interpretation?
- Is there any evidence or sources to back up the claims being made in the content?
- Does the content seem designed to manipulate or mislead the reader in some way?

---

---

---

---

---

### 6. False context

False context is information that is genuinely true, but shared with false contextual information, which can mislead and harm readers. For example, this could be using a photograph from the past and pretending that it is a recent one.

#### EXAMPLES



Just saw this picture of looters. They're all from [specific slum]! Disgusting!

A real image of looters is shared, but the caption falsely claims all the looters are from a particular slum, promoting prejudice.





Here's a video of a riot against refugees in Sudan. I can't believe this is happening!



A genuine video of a protest is re-shared with a claim that it's a violent riot against refugees, stirring up hostility. The video is in fact a protest from another context in a neighboring country.



**Learn to think critically**

- Is the information presented in a way that is designed to create a particular narrative or agenda?
- Is there any evidence that the information has been taken out of context or manipulated in some way?
- Are there any red flags that suggest the content may be misleading or manipulated?
- Is there any additional information or context that could help you better understand the situation?
- Is there any conflicting information from other sources that contradicts the claims being made in the content?

---

---

---

---

---

---

---

---

---

---

### 7. Manipulated content

Manipulated content is when genuine information is modified and distorted to mislead the users. For example, it could be a photo of someone in front of a house, with a headline that says "Look at this politician's gigantic palace!" when actually the house belongs to someone else.

#### EXAMPLES

During an election campaign, a video of a politician's speech was slowed down to make them appear unfit, drunk and unprofessional. It was manipulated and shared widely.

In 2018, a video of a reporter pushing away an intern who was trying to grab his microphone was widely shared. However, the video had been accelerated to appear more aggressive than it originally was.



#### Learn to think critically

- Has the content been digitally altered or manipulated in some way?
- Is the source of the content biased or intentionally misleading?
- Is there conflicting information from other sources that contradicts the claims being made in the content?
- Have I asked questions and sought out additional information to help me better understand the situation?
- Have I used reverse image search tools to see if the image has been used elsewhere or to find the original image?

---

---

---

---

---

---

## QUIZ 1

### Check your learning journey!

1

#### What is Dangerous Speech?

- a. A type of hate speech
- b. Communication that may help catalyze mass violence
- c. An exaggerated form of satire
- d. A form of disinformation

2

#### Which type of information refers to the spreading of false information without realizing it?

- a. Mal-information
- b. Disinformation
- c. Hate speech
- d. Misinformation

3

#### How does satire or parody contribute to the spread of mis-disinformation?

- a. By making outrageous demands
- b. By misleading use of information to frame an issue or individual
- c. By having potential to fool and mislead people
- d. By spreading false content, designed to harm and deceive

4

#### How can misleading content be identified?

- a. The tone of the post is humorous or sarcastic
- b. The headline seems too outrageous or exaggerated to be true
- c. The headline is sensational but not supported by the actual content of the article or post
- d. The source of the post is a well-known satire website

5

#### What is an example of imposter content?

- a. Social media post manipulates statistics about violence caused by RSF
- b. A fake account pretending to represent SAF releases statements endorsing violence
- c. A headline about a plan to exterminate ethnic minorities in Sudan
- d. A fabricated video circulates, falsely showing a politician from FFC praising the RSF's violent actions

#### ANSWERS

- 1) b, 2) d, 3) c, 4) c, 5) b

### 3. Why does it spread?



#### 1. The messages are simple

The easier a piece of information is to process, the more likely it is to be believed as true. Simple, clear, and repeated messages can feel more **familiar** and are easier to understand, which **increases their acceptance** and spread.

- **In reality, the truth is rarely as simple** as mis- and disinformation makes it appear to be. This information often connects to a bigger story.

"The Pink People are bringing problems to Sudan—just look at the violence they've caused."

"All we have to do is get rid of the Purple People and Sudan will be happy and peaceful at last. "

- **People tend to remember good stories**, long after they forget who told them the story. This can be dangerous – a good story that someone unreputable tells will likely be remembered better than a less compelling story told by someone that is reputable. The better the story and the more coherent it is, the more likely we are to remember (and potentially spread) it.



#### 2. The messages are repeated

When we see or hear a piece of information multiple times, we're more likely to believe it. This is known as the **"illusory truth effect"**. By repeating a lie, it can start to feel true.

- **Repeatedly hearing information can make us more likely to remember it— even** if it's not true. Repetitive information also starts to sound familiar, and because we may hear it from different sources, **we may think that the reason it's being repeated is because it's true**. After all, if other people believe it, why shouldn't I?

"I hear similar versions of the following statements four times over four weeks: The Purple people are the reason why our children have no food and famines have ravaged our community."

"I heard that the Purple people purposefully started this war to make us starve."



### 3. The messages appeal to our emotions

Emotionally charged content, particularly content that incites fear, anger, or outrage, is more likely to be shared. People are drawn to content that validates their emotions and viewpoints and are more likely to share such content.

- Mis- and disinformation often preys on people’s fears, disgust, and anger (using words like ‘evil’ or ‘punish’) to convince them to believe certain information.<sup>1</sup> We remember how things make us feel.
- When we’re filled with fear and other high-powered emotions, we might not be able to see that some information is not true, and that the author or source is not credible. Especially during periods of intense conflict, we can be more susceptible to disinformation **that preys on our already high levels of fear.**

#### Did you know ?

The illusory truth effect is a cognitive bias that causes us to believe information to be correct after repeated exposure. Essentially, if we hear something repeatedly, our mind starts to accept it as true, regardless of its accuracy. This effect can occur even when we initially recognized the information as false!

"Just found out a beloved local bookshop is closing because a big chain is opening next door. Heartbreaking! Stand up for small businesses!"

"Infuriating! Greedy landlords kicking out a sweet old lady who's lived here for 30 years! They are kicking her out because of her RELIGION! And now she will be alone in the cold street. Sign this petition with your email now!"

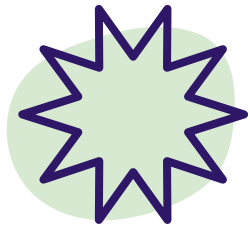
"Raging! Our children's future is being sold by the uncaring city council slashing school budgets! Demand better!"



### 4. We are all influenced by our peers and we experience conformity bias:

People tend to conform to the behaviors of groups or individuals around them. So, if it seems like everyone in your network is sharing a piece of information, you may be more inclined to do the same.

1 - Brady, W. J., Gantman, A. P. & Van Bavel, J. J. Attentional capture helps explain why moral and emotional content go viral. J. Exp. Psychol. Gen. 149, 746–756 (2020).



**5. The messages are sensational and crafted to catch our attention**

Dramatic or sensationalized content grabs attention. The more shocking or outrageous, the more likely it is to be clicked on, read, and shared.

"Heartbreaking! Popular charity organization scams millions from donors. Check if you were affected!"

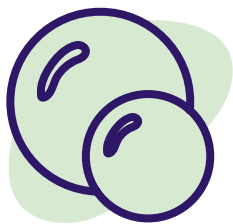
"Outrageous! Local Police caught hiding crime statistics. Your safety at risk!"

"Alert! New virus found in our city's water supply. Check your water sources now!"



**6. We often experience confirmation bias**

People have a natural inclination to seek, interpret, and remember information that confirms their pre-existing beliefs and ignore or discount information that contradicts them. This makes us more likely to accept and spread information that aligns with our worldview, regardless of its factual accuracy.



**7. We live in our bubble and experience the Echo Chamber Effect**

Social media algorithms often show us content similar to what we've liked or shared before, creating an "echo chamber" where we're more likely to encounter information that agrees with our views. This can amplify misinformation and make it spread more rapidly within certain communities.

## QUIZ 2

### Check your progress!

- 1**      **What percentage of social media users reshare or comment on posts without reading the article content?**
  - a. 30-40%
  - b. 60-70%
  - c. 80-90%
  - d. 10-20%
  
- 2**      **Simple, clear, and repeated messages are likely to be believed as true because:**
  - a. They are always accurate
  - b. They feel more familiar and are easier to understand
  - c. They contain sensational content
  - d. They confirm pre-existing beliefs
  
- 3**      **The "illusory truth effect" refers to the phenomenon where:**
  - a. Information repeated multiple times starts to feel true, regardless of its accuracy
  - b. People tend to believe information that confirms their pre-existing beliefs
  - c. Social media algorithms show us similar content we've liked or shared before
  - d. Information that incites fear, anger, or outrage is more likely to be shared
  
- 4**      **Emotionally charged content is more likely to be shared because:**
  - a. It confirms pre-existing beliefs
  - b. It is sensational and attention-grabbing
  - c. It validates users' emotions and viewpoints
  - d. It is easy to process
  
- 5**      **What does conformity bias suggest about information sharing behavior on social media?**
  - a. People tend to conform to the behaviors of groups or individuals around them
  - b. People only share content that is easy to process
  - c. People are more likely to share information that they have encountered multiple times
  - d. People are drawn to content that validates their emotions and viewpoints

**6**

### **Why is sensationalized content more likely to be shared?**

- a. It is always accurate
- b. It is repeated multiple times
- c. It grabs attention due to its shocking or outrageous nature
- d. It confirms pre-existing beliefs

**7**

### **What is confirmation bias in the context of information consumption and sharing?**

- a. The inclination to seek, interpret, and remember information that confirms pre-existing beliefs
- b. The phenomenon of believing repeated information to be true, regardless of its accuracy
- c. The tendency to share sensationalized content
- d. The habit of sharing content without reading it

**8**

### **How does the "Echo Chamber Effect" on social media contribute to the spread of misinformation?**

- a. It ensures that sensational content is shared more often
- b. It leads to the belief that repeated information is true, regardless of its accuracy
- c. It amplifies misinformation within certain communities by showing similar content to what users have liked or shared before
- d. It encourages the sharing of content without reading it

### **ANSWERS**

1) b, 2) b, 3) a, 4) c, 5) a, 6) c, 7) a, 8) c



## 4. What is Hate Speech?

### What is hate speech?

The United Nations defines hate speech as "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor."<sup>1</sup>

### Hate Speech can be private or public.

Private Hate Speech can spread via (virtual or in-person) conversations. Eg: Friends exchange racist jokes over afternoon coffee.

In public, it can spread via blogs, news reports, images, videos, songs<sup>2</sup>. Eg: An individual posts a video in which they and a group of their friends use offensive nicknames to describe certain tribes.

### What counts as Hate Speech?<sup>3</sup> and How can we recognize it?

#### 1. ATTACKS ON MINORITIES

Speech that attacks people based on their age, disability, ethnicity, religious affiliation.

**Example 1:** A video uploaded on a social media platform shows a speaker saying, "The Orange people are the cause of all our economic troubles."

**Example 2:** A popular podcast episode states, "Purple people are not compatible with our culture and should not be allowed to live among us."

#### 2. DENIAL OF HUMAN RIGHTS

Speech that advocates against rights to, for example, their language and traditions, basic standards of living, own property, religion and beliefs, and have basic legal protections.

**Example 1:** An online article advocates, "Pink people shouldn't be allowed to practice their religious rituals in our country."

**Example 2:** A public forum post argues, "Orange people should not be given the same legal protections as us because they don't contribute as much to society."

1 - United Nations, United Nations Strategy and Plan of Action on Hate Speech – Detailed Guidance on Implementation for United Nations Field Presences (United Nations, 2020).

2 - Miller-Idriss, Cynthia (2022). Hate in the Homeland: The New Global Far Right. Princeton, NJ: Princeton University Press.

3 - Jana Papcunová, Marcel Martoncik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogánová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovic. 2021. Hate speech operationalization: a preliminary examination of hatespeech indicators and their structure, Complex & Intelligent Systems.

### 3. PROMOTING VIOLENT BEHAVIOR

Encouraging terrorism, attacks and killing individuals or harming places in which certain groups may congregate (i.e. religious facilities).

**Example:** A user comments under a news article, "If Red people keep coming here, they are asking for trouble."

### 4. PROBLEMATIC HASHTAGS AND NICKNAMES

Speech that may treat people as less than or use derogatory metaphors.

**Example:** Users spread the hashtag #LazyYellow on Twitter, associating the Yellow people with laziness.

### 5. ATTACKS ON AN INDIVIDUAL'S CHARACTER

Speech that undermines an individual such as their worth, integrity, intelligence, or trustworthiness. Speakers may accuse individuals of lying, ignorance, or stupidity.

**Example:** An anonymous blog post attacks a well-known Purple person, a humanitarian and activist. The post reads: "**She only cares about her image, not about helping others. She's a liar, just like all Purple people.**"

**Example:** A public figure commented on a TV show, "**That Orange politician is untrustworthy. His lack of intelligence is proof that he's unfit for office. This is typical of Orange people—they're all ignorant and duplicitous.**"

### 6. NEGATIVE STEREOTYPES

Speech that reinforces offensive or demeaning traits of a group to which a person belongs to claim that the targeted group is inferior to other groups.

**Example:** An online comic strip consistently portrays **Pink people as lazy and unintelligent**. The narrative reinforces a stereotype that all Pink people lack motivation and intellectual capacity.

**Example:** A popular vlogger posts a video saying, "**Don't expect a Purple person to be good at sports. They're all bookworms who wouldn't know a basketball from a baseball.**"

### 7. AMBIGUOUS STATEMENTS AND IRONY

Speakers may use irony, sarcasm, Hate Speech memes, or talk jokingly as a way to spread Hate Speech without retribution. For example, they may mock victims of hate crimes.

**Example:** In an online forum, a user posts a seemingly humorous meme showing a Purple person slipping on a banana peel. The caption reads, "**Purple people can't even walk straight.**"

The message, while presented as a joke, perpetuates harmful stereotypes.

### 8. MANIPULATIVE TEXT

Speech that attempts to misinterpret the truth to fool others, like denying that historical events occurred.

**Example:** A widely-shared social media post states, "**There were never any concentration camps for Purple people. They made it all up for sympathy.**" This is a dangerous denial of historical atrocities committed against the Purple people.

**Example:** An online article claims, "**The Orange people weren't actually native to this land. They migrated here later, so they don't deserve any special rights.**" This false narrative aims to undermine the rights of indigenous Orange people.

### 9. SLURS AND VULGARISMS

Verbal or non-verbal attacks and insulting labels based on their race, ethnicity, religion, etc.

### 10. SEXISM

Spreading or justifying hatred based on an individual's sex, gender, or sexual preference; humiliating an individual to destroy their reputation and make them feel vulnerable.

**Example 1:** A user posts, "Orange people should stick to their traditional roles and leave leadership to Purple people."

**Example 2:** A viral post suggests that Purple people cannot be good engineers because they are too emotional.

**Categorization of Hate Speech: learn to recognize the early signs and the escalation**

Hate speech isn't always as explicit as we imagine it to be. It can often be misunderstood as absolute and extreme, thus justifying censorship or monitoring. However, this approach overlooks the subtler but steadily escalating forms of hate speech that can eventually turn harmful. Identifying these milder forms early on is essential in order to preempt their potential to escalate. To this end, we can leverage a scale created by Babak Bahador<sup>1</sup>, allowing us to detect and address hate speech in its early stages, before it spirals into tangible harm.

| Colour | Title                              | Description                                                                                                                                                                                      | Examples                                                     |
|--------|------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
| 6      | <b>Death</b>                       | Rhetoric includes literal killing by group. Responses include the literal death/elimination of a group.                                                                                          | Killed, annihilate, destroy                                  |
| 5      | <b>Violence</b>                    | Rhetoric includes infliction of physical harm or metaphoric/aspirational physical harm or death. Responses include calls for literal violence or metaphoric/aspirational physical harm or death. | Punched, raped, starved, torturing, mugging                  |
| 4      | <b>Demonizing and Dehumanizing</b> | Rhetoric includes subhuman and superhuman characteristics. There are no responses for #4.                                                                                                        | Rat, monkey, Nazi, demon, cancer, monster                    |
| 3      | <b>Negative Character</b>          | Rhetoric includes nonviolent characterizations and insults. There are no responses for #3.                                                                                                       | Stupid, thief, aggressor, fake, crazy                        |
| 2      | <b>Negative Actions</b>            | Rhetoric includes negative nonviolent actions associated with the group. Responses include nonviolent actions including metaphors.                                                               | Threatened, stole, outrageous, act, poor treatment, alienate |
| 1      | <b>Disagreement</b>                | Rhetoric includes disagreeing at the idea/belief level. Responses include challenging claims, ideas, beliefs, or trying to change their view.                                                    | False, incorrect, wrong, challenge, persuade, change minds   |

1 - <https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/>

**Did you know ?**

Social media algorithms are the reason why **two users will NOT see exactly the same social content, even if they follow all the same accounts?** This is because algorithms use a set of rules and signals to automatically rank content on a social platform based on how likely each individual social media user is to like it and interact with it. So, even if two users follow the same accounts, their past behavior and interactions with the platform will influence what content is shown to them.

## 5. Social media: the fuel and the fight

**Social media, for all its worldwide influence, has a uniquely important role in the context of Sudan's Dangerous Speech. In essence, it serves a dual function, acting both as the fuel and the fight against Dangerous Speech. Here's why:**

### 1. SOCIAL MEDIA IS AN AMPLIFIER OF DANGEROUS SPEECH

Dangerous Speech, online or offline, isn't confined to a single platform or space. Social media platforms, due to their rapid spread of information, provide a fertile ground for Dangerous Speech to thrive and amplify. Dangerous Speech can find its way to individuals who aren't even active on social media, as they get exposed through their offline social networks and community interactions.

### 2. CONNECTIVITY IS NOT BINARY—IT'S NOT AN ALL-OR-NOTHING SITUATION

Addressing Dangerous Speech is a complex, multifaceted task due to the varying degrees of online content exposure among people. Even if some individuals lack personal access to smartphones, they might still be exposed to content by using devices belonging to their friends or family. Thus, the fight against Dangerous Speech extends beyond direct users, reaching those with indirect exposure as well.

### 3. RESPONDING TO DANGEROUS SPEECH / FINDING SOLUTIONS NEEDS TO HAPPEN ONLINE AND OFFLINE AS WELL

Given its omnipresence, effectively addressing Dangerous Speech calls for a multi-pronged approach that involves combating the issue both online and offline. This includes raising awareness and promoting counter speech on social media, alongside engagement with community leaders, organizations, and public spaces to foster positive social norms and discourage harmful speech.

### Did you know ?

Social media platforms, with their algorithm-driven feeds, often create "echo chambers". This term refers to the phenomenon where **users are primarily exposed to content that aligns with their beliefs**, further entrenching existing views and biases, and potentially amplifying the spread of Dangerous Speech.

### 4. SOCIAL MEDIA CAN ALSO BE A POWERFUL TOOL TO FIGHT DANGEROUS SPEECH

Despite its role in propagating Dangerous Speech, social media is also an invaluable tool in the fight against it. Its capacity to **reach wide audiences** and **spread messages quickly** can be leveraged to challenge and change harmful social norms.

Influencers, activists, and everyday users alike have the power to use these platforms to break down social barriers and promote tolerance and respect for diversity. Combatting Dangerous Speech on social media contributes to societal change by making **positive behaviors visible, accepted, and celebrated**.

### Did you know ?

#### The Power of the 'Share' and "like" Buttons:

1. **They feed the algorithms:** social media algorithms are designed to prioritize content that receives high engagement (likes, shares, comments) because it's perceived as interesting or valuable to users. So, when a post gets many likes or shares, the algorithm will promote it to more people, making it more visible across the platform. This can result in the rapid and widespread dissemination of a piece of content, including Dangerous Speech.
2. **They influence other users:** The 'Like' and 'Share' buttons also serve as indicators of social proof, a psychological and social phenomenon where people's attitudes, beliefs, and actions are influenced by others. Essentially, if a post has many likes or shares, people may perceive it as more credible or important, regardless of its actual accuracy or value. This can exacerbate the spread of Dangerous Speech, as people might be more likely to share or believe such content if they see others engaging with it.

2

# Chapter 2

## Learn how to respond to dangerous speech

### Did you know ?

We tend to underestimate how influential we are.<sup>1</sup> People often underestimate their social networks' size and their own significance within these circles. It's also easy to overlook the breadth of influence we have over various individuals. Not only do we influence our friends' opinions, but we can also impact strangers. Our influence extends across ages too - it's not just limited to our peers, but also younger and older folks.

### Key message:

**While it might seem instinctive to combat Dangerous Speech with counter-arguments, research and experience suggest that such an approach can often be counter-productive.** Confronting trolls or perpetrators head-on can result in wasted energy and an entrenched opposition. It's a phenomenon known as the 'backfire effect,' where direct attempts to correct misinformation can, paradoxically, reinforce the false belief. Instead, this section provides evidence-based strategies to pre-emptively challenge Dangerous Speech (known as pre-bunking), debunk effectively when Dangerous Speech is already circulating, and counterspeak with a thoughtful understanding of your audience. We will delve into the common cognitive biases that can influence our susceptibility to Dangerous Speech and guide you on how to address these effectively. By employing these approaches, we can ensure our efforts have the maximum impact in mitigating the spread and influence of Dangerous Speech.

### 1. Arguing is counterproductive and rarely works

The intuitive, obvious way to answer DS can be counterproductive. If you see a piece of DS and start answering with arguments, it is likely that you will waste your energy and that you will have a hard time convincing the trolls or perpetrators. In this section, we will teach you to recognize the most common cognitive biases and answer them, to pre-bunk to prevent the spread of DS, to debunk effectively, and to counterspeak with the audience in mind.

<sup>1</sup> - Bohns, V. K. (2021). You have more influence than you think: How we underestimate our power of persuasion and why it matters. W. W. Norton & Company.

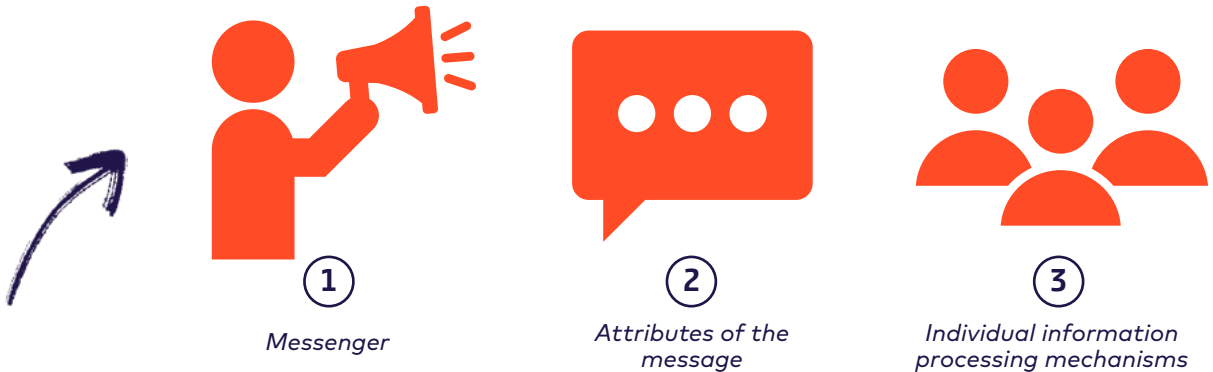


## 2. Most common cognitive biases and how to answer them


Understanding the detrimental effects of Dangerous Speech is one thing, but altering our behavior based on this knowledge is another. Even if we understand how misinformation spreads, this realization alone may not be enough to change our actions.

**Simply put, knowing the facts doesn't automatically mean we accept or believe them, and people's reactions to the same information can vary considerably.**

Behavioral science highlights how our **acceptance** and **dissemination** of information hinges on (1) **who communicates it**, (2) **the specific attributes of the message**, and our (3) **individual information-processing mechanisms**.



### 1. Why does everyone keep spreading Dangerous Speech even when they know it's wrong?

 Social norms: This is when we start to act a certain way because others in our community are doing so. If you are part of a group that uses Dangerous Speech, you may feel pressure to use it to belong. If we see our friends, families, and others around us use Dangerous Speech and Hate Speech, we may feel like it's acceptable and do so as well.

#### EXAMPLE

Most Pink People in your community do not like Purple People and often use inflammatory terms like "PP" when referring to them and how they've hurt Pink People. As a Pink Person, you've met very peaceful Purple People and are even friends with Yellow people members. However, you feel embarrassed to go against your friends and family, and you don't want to be ostracized, so you start using these terms as well.

YOU HAVE MORE INFLUENCE THAN YOU THINK



**How can you combat it? Change the norm. You have more influence than you think! If not you, then who? If not now, then when?**

Remember that there may be many more people like you who do not support Dangerous Speech but who are also silent. The first step to changing the norm is to break the silence so that those people can follow. Together, you can create a critical mass.

### 2. What makes information more credible when it comes from a friend, even if I'm saying the exact same thing?



**Messengers:** We are greatly affected by who tells us information. We are often more receptive to information if it comes from individuals that are **similar to us**, or from people that we perceive as **attractive** or **powerful** (both official and non-official authority figures, like religious leaders)<sup>1</sup>. **The more credible or likeable the person is, the more persuasive the message becomes.**



#### EXAMPLE

Imagine this scenario. A rumor is circulating around about a new, deadly snake species discovered in the park. Fatima decides to verify this claim. After hours of research through trustworthy news sites and speaking with a local scientist, she figures that the snake rare but it is not dangerous.

Eager to stop the rumor, Fatima shares her findings with her acquaintance, Ibrahim, providing all the necessary sources and facts to support her claim. However, Ibrahim also hears about the snake rumor **from his boss**, who is convinced the snake is lethal.

**Even though Fatima provides Ibrahim with the same factual information, Ibrahim trusts his male boss's version more.** Despite her verified information, Fatima's gender seems to influence Ibrahim's reception of her information. **It's not just about what is said, but also about who says it.**

<sup>1</sup>- Briñol, P. & Petty, R. E. Source factors in persuasion: a self-validation approach. Eur. Rev. Soc. Psychol. 20, 49–96 (2009).

### Did you know ?

We tend to give preferential treatment for groups to which we belong. This is called in-group favoritism. We are also more at risk of negatively perceiving or treating outgroups, the groups to which we do not belong.



### How can you combat it?

Fighting the messenger effect is difficult but not impossible. One powerful strategy is to leverage the influence of trusted group members who share common identifiers, such as religious or cultural beliefs. You can work with them to ensure they spread the right messages, fueled by science.

### 3. Why are different groups so hateful towards one another?



**In-group bias:** An 'in-group' is any identity that we have that is tied to belongingness to a group. This could be your membership to a sports team, your ethnicity, religion, political beliefs, or other 'groups'. When we belong to a group, we can often give preferential treatment to others in that group, and treat those who are not in our group (the 'out-group') less positively or support violence against them. Even if we don't believe certain ethnic groups are bad, we may treat them poorly because we are afraid of being ostracized by people that are part of our group.

### EXAMPLE

Members of the Orange group are sitting together playing cards and talking about the rising unemployment rate. While they are doing everything they can to provide for their families, they grumble about how lazy Purples are. A Purple member had the audacity to ask an Orange member for a job recently - why would an Orange give a job to someone who's clearly unambitious?



### How can you combat it?

1. Engage in perspective giving to help people understand each other's experiences. Research shows that putting yourself in someone's shoes might not be the most effective way to build empathy. Instead, helping people gain perspective by simply telling them how others feel can be effective.
2. Engage in analog perspective taking: reflect about a time where something similar happened to you, and transport that feeling to the situation currently at play.

3. Encourage shared identities: remind individuals of their similarities, such their shared love of football, their identities as students, farmers, colleagues, fathers, brothers, or sisters, and their shared favorite food.
4. Contact: do your best to get to know people from your outgroups, share activities, information, learn from them and see the world from their viewpoint. Research suggests that meaningful interactions with people from outgroups can help bring a fresh perspective on these biases.



### Prompts to cultivate empathy:

- Imagine them as a member of your family: How would your understanding and tolerance of their situation change if they were your sibling or parent?
- If they were your best friend dealing with this situation, how would you feel, and what would you do to support them?
- Think about a time where something similar happened to you ? How did that make you feel ? How similar is it to what people are experiencing in this situation?
- Visualize people’s struggles and successes: how have their challenges and victories shaped who they are, and how can this understanding deepen your empathy for them?
- What common ground can you find with them? Despite your differences, what similarities can you identify?

---

---

---

---

---

---

---

---

---

---

#### 4. Why do people believe rumors?



**Illusory truth effect:** We tend to believe information if we've heard it multiple times, even if it's false. If people continue to repeat mis or dis-information, that information starts to be believed as fact.

**Confirmation bias:** People will tend to look for and remember information that supports their worldview. They may forget when a member of the outgroup is kind and will vividly remember a negative interaction with a member of the outgroup.

#### EXAMPLE

**Example:** Imagine you constantly hear a rumor on social media that says "Purple people always take more than their fair share of resources". You hear it so often, it's practically echoing in your ears - this is the **illusory truth effect** in action, and even though it's baseless, the **repetition makes it seem like a fact**.

Furthermore, **you've always felt that the Purple people in your town have larger houses**, so this rumor seems **to confirm what you already thought** - this is your confirmation bias. You easily forget that you know some Orange people with big houses too and that the size of a house doesn't determine how much one consumes. But because the rumor aligns with your preconceived notions, you are more likely to believe and propagate it, spreading the disinformation further.



#### How can you combat it?

Avoid the repetition of the rumor and reinforce the correct information instead. Show counter-examples in a captivating manner that appeals to people's emotions (positive stories of human connection) or back trusted sources that challenge these rumors (scientists, experts, community leaders).

### 5. Why do people believe misinformation?



**Emotions:** Initial feelings like worry or anxiety can make us less likely to critically evaluate information. This can make us more likely to believe misinformation and Dangerous Speech/Hate Speech.<sup>1</sup>

**System 1:** When we are rushed, distracted or busy, we can dedicate less time to carefully and critically examine the information received.

#### EXAMPLE

**Example:** Imagine you are having a hectic day at work with back-to-back meetings and an overflowing inbox. In between, you receive a message on your phone from a friend saying, "**Purple people are planning a protest in our neighborhood tonight. They're going to disrupt the peace! Go and grab all the medicines from the nearest pharmacy, we never know.**" Given the stress you're under, you don't have the time or mental bandwidth to question this message. You might go and hoard all the medicine from the nearest pharmacies, even though you don't need it and others might need it more than you.

The initial wave of anxiety brought on by the message discourages you from scrutinizing it further. This is your System 1, the automatic, fast-thinking part of your brain, taking over. As a result, you're more likely to believe this misinformation and even share it with others, thus contributing to the spread of Dangerous Speech.



#### How can you combat it?

Encourage people to **pause and rethink**. Ask them how they know the information is true or false. If we take time to process and evaluate the information, we're less likely to believe mis- and dis-information.<sup>2</sup>

1 - Brashier, N. M. & Marsh, E. J. Judging truth. *Annu. Rev. Psychol.* 71, 499–515 (2020).

2 - Bago, B., Rand, D. G. & Pennycook, G. Fake news, fast and slow: deliberation reduces belief in false (but not true) news headlines. *J. Exp. Psychol. Gen.* 149, 1608–1613 (2020)

## 6. Why don't people stop believing misinformation after they know it's not true?



**Continued Influence effect:** This cognitive bias occurs when individuals continue to **rely on incorrect information to guide their thinking and decision-making**, even when they've been exposed to accurate information that contradicts the original misinformation. **Misinformation can often continue to influence people's thinking even after they receive a correction and accept it as true.**

### EXAMPLE

Several years ago, rumors circulated that individuals in the Purple region were responsible for a bus crash that killed dozens of Sudanese civilians from the purple region. Later, it was deemed that the bus crash was caused by a storm, and there was no foul play. However, even after the rumor had been shown untrue, it continued to spread throughout communities. Years later, many residents still believed the rumor.



### How can you combat it?

The correct information needs to be strengthened through repetition. Be careful not to inadvertently repeat the myth when you are correcting it.

**Overkill Backfire Effect:** Myths are often simple and sound like the most "obvious" scenario. They don't require much deliberate thinking to be remembered and likely build on pre-existing stereotypes. The reality is actually much more complex, and refutations using facts and evidence need more mental capacity to be understood. Refutations are also less appealing because they are countering long-standing stereotypes and preconceived notions.

**How can you combat it?** Keep facts short and easy to remember. Use visual aids like a figure or a chart or a picture to convey your message. Use fewer messages to refute the myth. Remember, less is more!<sup>1</sup>

<sup>1</sup> - Lewandowsky, Stephan & Ecker, Ullrich & Seifert, Colleen & Schwarz, Norbert & Cook, John. (2012). Misinformation and Its Correction Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*. 13. 106-131. 10.1177/1529100612451018.





### Did you know ?

Sudanese youth rely heavily on social media for information, with over **90% considering the information they receive to be either very or somewhat reliable.**



## 3. Use pre-bunking techniques to prevent Dangerous Speech from spreading and harming others



Pre-bunking, is the fusion of 'preemptive' and 'debunking', involves using strategic communication to provide individuals with information that can help them recognize and resist misinformation before they encounter it. It's a powerful tool in fighting dangerous speech and misinformation. Here's how to use pre-bunking techniques effectively:

### 1. EDUCATE ON THE SPREAD OF MISINFORMATION:

The idea: Familiarize your audience with common misinformation tactics.

You can share articles or create content that explains how fake news and dangerous speech are created and spread.

#### EXAMPLE

**Did you know? Sometimes, misinformation uses sensational headlines to grab your attention. Always look beyond the headline before sharing!**

### 2. EXPOSE COMMON SOURCES OF MISINFORMATION

The idea: Reveal prevalent sources of misinformation.

You can keep track of websites, channels, or accounts known for sharing unverified information and educate your audience about them.

#### EXAMPLE

**Watch out for 'XYZ News'; they've been known to share unchecked stories!**

### 3. HIGHLIGHT FACT-CHECKING RESOURCES

The idea: Promote the habit of fact-checking.

You can share links to credible fact-checking websites and demonstrate how to use them.

#### EXAMPLE

**Not sure if a story is true? Use fact-checking websites like 'FactCheck.org' to verify!**

### 4. PROMOTE CRITICAL THINKING

The idea: Encourage your audience to question and analyze the information they consume.

You can share guides or tips on critical thinking and ask probing questions about shared information.

#### EXAMPLE

**Before sharing that post, ask yourself: who benefits from this information? Is there any evidence to support it?**

### 5. PREEMPT POTENTIAL MISINFORMATION

The idea: When you're aware of an event that might be a target for misinformation, provide accurate information beforehand.

You can be proactive in sharing verified information on hot topics.

#### EXAMPLE

**With the upcoming election, remember to verify your news from reliable sources and don't fall prey to sensational claims without evidence!**



### Tips to get people's attention when you debunk:

1. Correct the misinformation quickly, the longer it circulates, the harder it is to correct.
2. Use visuals to help people understand complex information faster.
3. Be clear and concise.
4. Be kind and respectful. You risk losing people's trust if you speak with emotions or if you are condescending.

## 4. De-bunk dangerous speech with evidence based techniques



Debunking is a reactive measure to fact-check and expose Dangerous Speech that has already spread. It's different from pre-bunking because it's about making a direct response to content that is already circulating, while pre-bunking is a proactive measure to prevent information from spreading. Here is how to do it effectively:

### 1. State why the information is not credible

If you have proof, state why the information is wrong. If you don't have proof, declare your skepticism about the information (sources unknown, missing context, location or time or important information, inciting violence).

Offer alternative facts and causal explanations.

### 2. Alert your audience to the problems in the dangerous speech

Expose the inaccuracies and lack of credibility in the harmful message. Break down the problematic aspects of the information, including its origin and intentionality. Highlight its aim to incite fear, stoke hatred, justify violence, and obscure the actions of the perpetrator.

### 3. Highlight your own values to build trust and credibility

Emphasize your personal principles: being impartial, relentlessly pursuing truth, and maintaining a willingness to scrutinize your own perspectives. Emphasizing personal principles is crucial as it builds **trust** and **credibility**, two key factors that can influence the acceptance of your counter-narrative.

### 4. Foster a new social norm

Promote a culture of thoughtful interaction by urging others to actively question and comment when they encounter questionable content. You can set the example to follow. Remember that you have more influence than you think!

## 5. Be an Active Bystander to Support Victims (online and offline)

An active bystander is someone who observes a problematic situation and **proactively takes steps** to intervene, aiming to **prevent escalation or disruption**. Unlike a standard bystander, who merely observes, an active bystander might speak out against offensive behavior, report the situation, or support the targeted individual. Their actions, essentially promoting a culture of respect and positivity, play a pivotal role in harm prevention.

When you witness a problematic situation online, you can be an active bystander following the 4Ds approach: **Distract, Delegate, Delay** or **Direct** approach.<sup>1</sup>



### Distract

Devise a diversion to break the ongoing harmful interaction. This could be by redirecting the conversation or introducing a new topic.

### Delegate

Don't hesitate to involve others. If the situation is escalating, report it to the relevant authorities or seek support from individuals in your community.

### Delay

If immediate action isn't possible or safe, you can respond later. This could be by checking in with the targeted individual after the incident or reporting the situation once you're able to do so.

### Direct

If it's safe and appropriate, address the situation directly. Speak out against the harmful behavior or comment. However, always ensure your own safety is not compromised when taking a direct approach.

<sup>1</sup> - Bystander Intervention: A Critical Step To Prevent Harassment (everfi.com)

### Tips for successful counterspeech:

1. **Warn the perpetrator of the consequences:** remind the speaker that their speech is harmful to others, and state the consequences it can have. Remind them that online communications are permanent and leave a footprint. Remind them that their family members or relations can see what they are posting.
2. **Labeling:** denounce the speech as hateful or dangerous and explain to the speaker why their speech is hateful. This might prevent them from repeating that mistake.
3. **Show empathy and affiliation<sup>19</sup>:** use a friendly, empathetic, peaceful tone to prevent further escalation.
4. **Use humor:** this can shift the dynamic of the conversation, de-escalate conflict, and draw more attention to your message.
5. **Use Images:** pictures, memes, animated gifs can make content go viral and tap into the spectators emotions.

## 6. Influence the spectators : the power of Counterspeech on social media

**Counterspeakers** are people who respond directly to online dangerous speech in an effort to improve discourse. They do this by addressing the spectators, who usually far outnumber people who post hateful content. Simply put: there are more spectators than there are perpetrators.

There are 4 main reasons to address the spectators with counterspeech:

1. Changing the spectator's views,
2. Recruiting new counterspeakers,
3. Strengthening norms against negative content among the audience,
4. Supporting those targeted by the Dangerous Speech.<sup>1</sup>

You may not be able to convince the perpetrators, sometimes called the "trolls", and that's okay. The most important is to counterspeak anyway, so you can convince the spectators. That should be your main goal. It will make your work more impactful.

Attention, humor is not condescending satire, but actual jokes as illustrated in the example below.<sup>2</sup>



1 - Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., ... & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50), e2116310118

2 - Considerations-for-Successful-Counterspeech.pdf (dangerousspeech.org) S. Benesch, D. Ruths, K. P. Dillon, H. M. Saleem, L. Wright, Considerations for Successful Counterspeech (Dangerous Speech Project, 2016).

3

# Chapter 3

Practice fighting Dangerous Speech with exercises & quizzes

## Exercise 1

Look at this facebook post, and answer the questions to the best of your ability:



Facebook User 01

Today at 2:04pm

Those Purple folks are taking our jobs and our livelihoods, benefiting from our hard-earned social benefits. Is there an end to this drain on our resources? Now, due to this strain, our cost of living has spiked by 8%. How am I supposed to put food on my table for my kids?



### REFLECT

1. Why is this post a form of dangerous speech?
2. What mechanism is this author using to make this message appealing ?
3. How would you respond to this post if you saw it on your feed?
4. Can you think of any ways to address this issue without directly engaging with the person who made the post?

---

---

---

---

---

---

---

---

## Exercise 2

Look at this twitter post, and answer the questions to the best of your ability:



**Twitter User 01**

@twitteruser01

Have you heard about the woman who complained about harassment at work? Well, she was practically begging for attention with her flamboyant style! Her claims are baseless and she’s only playing the victim for sympathy. Don’t be fooled. This is just another ploy by the feminists to tarnish the reputation of good men. She should learn to keep a low profile if she doesn’t want unwelcome attention. Time for the ‘MeToo’ madness to end!



### REFLECT

1. What would counterspeech look like that incorporated empathy?
2. What would counterspeech look like that incorporated humor?
3. What would counterspeech look like that incorporated a warning of consequences?
4. Can you think of any local social media influencers that could help spread counterspeech? What about other credible messengers?

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---



### Exercise 3

Look at this Twitter post, and answer the questions to the best of your ability:



**Twitter User 01**

@twitteruser01

This country is suffering! Economic crisis, no food, no healthcare, poverty going up, and children barely going to school! People can't afford the basic necessities anymore! Want to know why? There are no jobs! And why is that? It's because we have so many Orange people that are willing to work for cheaper who are taking YOUR jobs. We are being replaced. Our way of life is being changed due to external agendas. We need to fight back, otherwise this country will become a ORANGE country!"



#### REFLECT

1. How might someone feel when they hear this?
2. What type of Dangerous Speech do you see?
3. Are there any behavioral insights that can help you understand why someone might believe or spread this sentiment?
4. How might you try to combat this Hate Speech/Dangerous Speech using behavioral science?

---

---

---

---

---

---

---

---

---

---

---

### Exercise 4

Look at this Twitter post, and answer the questions to the best of your ability:

You see an online Twitter post by a well-known government official that warns residents against associating with certain tribes, because that’s how COVID-19 spreads. You believe every effort should be done to prevent further COVID infections and you have previously shared information about COVID warnings.

*The messenger and you have the same need to slow down the pandemic and keep people safe, except that they are using false information.*



#### REFLECT

1. What steps would you take before engaging with or sharing this post?
2. How would you teach others to think critically about this kind of content in the future?



A light blue rectangular area containing ten horizontal lines for writing answers to the reflection questions.

### Exercise 5

Look at this twitter post, and answer the questions to the best of your ability:



Twitter User 01  
@twitteruser01

As a Orange person, I only hire Purple people in my detergent factory. I get to try all my new products on them first hahaha



#### REFLECT

1. How do you feel about this information?
2. What behavioral concepts does this statement draw on?
3. What positive examples could you use to counter this?

---

---

---

---

---

---

---

---

---

---

---

---

---


---


---

---

**Exercise 6**

Look at this picture shared on Whatsapp and Facebook, and answer the questions to the best of your ability:

 **WhatsApp User 01**  
Today at 6:57pm



*IF NOT NOW, THEN WHEN?*

The photo above was taken at the football stadium after the Purple people left. The filth says it all. Act now before they do this to your own country. Sign this petition to #kickpurpleout



**REFLECT**

1. How is the messenger using dangerous speech?
2. How could you verify the image/picture?
3. How can you teach others to verify images and their context before forwarding to their networks?

---



---



---



---



## QUIZ 1

### Practice fighting Dangerous Speech

- 1** **What is debunking in the context of countering Dangerous Speech?**
  - a. A proactive measure to prevent misinformation from spreading
  - b. A reactive measure to fact-check and expose Dangerous Speech that has already spread
  - c. An attempt to provide alternative facts and causal explanations to disputed information
  - d. Both b and c
  
- 2** **Which of the following is not a part of an effective debunking strategy?**
  - a. Stating why the dangerous information is not credible
  - b. Alerting the audience to the problems in the dangerous speech
  - c. Highlighting your personal biases to build trust
  - d. Fostering a new social norm
  
- 3** **What is the role of an active bystander in preventing the spread of Dangerous Speech?**
  - a. Observing a problematic situation passively
  - b. Intervening in a problematic situation to prevent escalation
  - c. Sharing misinformation to make it more visible
  - d. None of the above
  
- 4** **Which of the following is not one of the 4Ds approach used by an active bystander?**
  - a. Distraction
  - b. Delegation
  - c. Delay
  - d. Destruction
  
- 5** **Counterspeakers address the spectators during online dangerous speech because:**
  - a. There are more spectators than perpetrators
  - b. It is easier to convince the spectators than the trolls
  - c. They can recruit new counterspeakers among the spectators
  - d. All of the above

**6**

**Which of the following is not a tip for successful counterspeech?**

- a. Using humor
- b. Ignoring the perpetrator
- c. Showing empathy and affiliation
- d. Using images

**7**

**When debunking Dangerous Speech, why is it important to highlight your own values?**

- a. To assert dominance over the conversation
- b. To build trust and credibility
- c. To emphasize your expertise in the subject
- d. None of the above

**8**

**In the context of being an active bystander, what does 'Delay' in the 4Ds approach refer to?**

- a. Ignoring the issue until it goes away
- b. Delaying your own reaction to provoke a response from others
- c. Responding to the situation later when immediate action isn't possible
- d. Waiting for someone else to intervene first

**9**

**What is the main aim of counterspeaking?**

- a. To convince the perpetrators to stop spreading dangerous speech
- b. To convince the spectators not to believe the dangerous speech
- c. To entertain the spectators with humorous counterspeech
- d. All of the above

**10**

**Which of the following is a technique not recommended when debunking Dangerous Speech?**

- a. Reacting with emotion or condescension
- b. Using visuals to help people understand complex information
- c. Correcting the misinformation quickly
- d. Being clear and concise

### ANSWERS

1) d, 2) c, 3) b, 4) d, 5) d, 6) b, 7) b, 8) c, 9) b, 10) a

## QUIZ 2

### Practice answering common cognitive biases

1

**Which of the following best defines social norms?**

- a. The tendency to remember information that supports one's worldview
- b. b) The tendency to behave similarly to how others are behaving
- c. c) The tendency to believe information that has been heard multiple times
- d. d) The way we behave based on our fear emotions

2

**How can one combat the effect of social norms?**

- a. By breaking the silence to show a new norm
- b. b) By ignoring the norm and keeping silent
- c. c) By conforming to the norm to avoid conflict
- d. d) By challenging the norm with aggression

3

**Why are we often more receptive to information from people we perceive as attractive or powerful?**

- a. Because of the messenger effect
- b. b) Because of the in-group bias
- c. c) Because of the illusory truth effect
- d. d) Because of the confirmation bias

4

**What is a powerful strategy to fight the messenger effect?**

- a. Ignoring messages from unlikeable people
- b. b) Discrediting the source of information
- c. c) Leveraging the influence of trusted group members who share common identifiers
- d. d) Spreading your own version of the message

5

**What is in-group favoritism?**

- a. Giving preferential treatment to groups to which we belong
- b. b) Treating all groups with the same respect
- c. c) Favoring groups based on the benefits they provide
- d. d) Favoring outgroups for the sake of diversity



**6**

### **What can one do to combat in-group bias?**

- a. Isolate themselves from other groups
- b. Actively adopt the perspective of others and encourage shared identities
- c. Avoid contact with people from outgroups
- d. Deny any bias and treat everyone the same

**7**

### **How can one combat the effect of social norms?**

- a. By breaking the silence to show a new norm
- b. By ignoring the norm and keeping silent
- c. By conforming to the norm to avoid conflict
- d. By challenging the norm with aggression

**8**

### **How can one combat the illusory truth effect?**

- a. By repeating the rumor
- b. By ignoring the rumor
- c. By reinforcing the correct information and avoiding the repetition of the rumor
- d. By creating a new rumor to divert attention

**9**

### **What role does emotion play in the belief of misinformation?**

- a. Emotions make us more rational and skeptical
- b. Emotions like worry or anxiety can make us less likely to critically evaluate information
- c. Emotions have no influence on our belief of misinformation
- d. Emotions help us verify the credibility of the information we receive

**10**

### **How can one combat the influence of emotions in the belief of misinformation?**

- a. By ignoring their emotions
- b. By allowing their emotions to guide their decision
- c. By encouraging people to pause and rethink the information they receive
- d. By reinforcing their emotions with more misinformation

**11****What is the Continued Influence effect?**

- a. The cognitive bias where individuals continue to rely on incorrect information even when they've been exposed to accurate information
- b. The cognitive bias where individuals are influenced by the group's opinions
- c. The cognitive bias where individuals believe information that they've heard multiple times
- d. The cognitive bias where individuals are influenced by the person delivering the information

**12****How can one combat the Continued Influence effect?**

- a. By repeating the correct information and avoiding repeating the myth
- b. By repeating the myth until it is proven false
- c. By avoiding all information about the subject
- d. By agreeing with the incorrect information

**13****What is the Overkill Backfire Effect?**

- a. It's when complex facts and evidence used in refutations are less appealing because they require more mental capacity to be understood
- b. It's when simple rumors are less appealing because they require less mental capacity to be understood
- c. It's when refutations are less appealing because they confirm long-standing stereotypes
- d. It's when refutations are more appealing because they challenge long-standing stereotypes

**14****How can one combat the Overkill Backfire Effect?**

- a. By making refutations complex and difficult to understand
- b. By making refutations longer and more detailed
- c. By using fewer messages to refute the myth, keeping facts short, and easy to remember
- d. By ignoring the myth and not providing any refutations

**15****What is confirmation bias?**

- a. A tendency to look for and remember information that contradicts our worldview
- b. A tendency to look for and remember information that supports our worldview
- c. A tendency to deny any information that supports our worldview
- d. A tendency to accept all information that comes our way

**16****How can one cultivate empathy to combat in-group bias?**

- a. By visualizing others' struggles and successes
- b. By ignoring others' perspectives
- c. By focusing solely on one's own experiences
- d. By maintaining the boundaries between in-groups and out-groups

**17****Why do people not stop believing misinformation even after knowing it's not true?**

- a. Because of the Continued Influence effect
- b. Because of the Overkill Backfire Effect
- c. Because they want to spread the misinformation
- d. Because they do not trust the source of the correct information

**18****How can people combat the effects of confirmation bias and the illusory truth effect?**

- a. By accepting all the information they come across
- b. By avoiding the repetition of the rumor and reinforcing the correct information
- c. By promoting the rumor to draw attention to it
- d. By ignoring all the information related to the rumor

**ANSWERS**

1) b, 2) a, 3) a, 4) c, 5) a, 6) b, 7) b, 8) c, 9) b, 10) c, 11) a, 12) a, 13) a, 14) c, 15) b, 16) a, 17) a, 18) b/d

4

# Chapter 4

## Templates for social media

### 1. Pre-bunking template: don't get overwhelmed, just follow the steps!

**1****Fact**

Start with the fact. Make it clear and sticky.

"Blue people are not responsible for the crimes that are so often being attributed to them."

**2****Warning**

Warn the audience about misinformation.

"Many posts are linking crime, theft, illegal practices to Blue people. We've seen a surge of those in the past few months."

**3****Fallacy**

Identify the biases, fallacies and technics employed.

"Linking Blue people to crime is meant to distract away from the real problem and to stir up collective fears without base."

**4****Critical Thinking**

Encourage critical thinking and promote reputable sources.

"If it's too simple or too shocking, maybe it's not true. Ask yourself these two questions before sharing." or "Next time, check the verified AFP website and use a reverse image search to check before you share."

**5****Fact**

Conclude with replacing the misinformation by the correct information.

"The rise in crime rate has actually not increased as compared to previous years according to national statistics."



## 2. De-bunking template don't get overwhelmed, just follow the steps!



1. State why it is not credible if you have proof or declare your skepticism about the information (sources unknown, missing context, location or time or important information, inciting violence)
2. Warn audiences about what makes it not credible (origin, intentionality- creating fear, fueling hate, justifying violence, hiding perpetrator)
3. Highlight your own values (unbiased, seeking only the truth, willing to be critical of your own views)
4. Foster a social norm by encouraging people to critically engage and comment when they have doubts about the credibility of the information to help others not fall prey to Dangerous Speech



### Fact

The stadium was not polluted by the blue people

### Evidence

A simple reverse search of the image shows that it was taken from elsewhere

### Warning

There is a lot of misinformation about what happened at the stadium

### Critical thinking

If it's too simple or too shocking, you can use the same reverse image search to check

### Values

We all value unbiased reporting and it's important we check facts before sharing

### Promote engagement

Participate in the comments section when you have these doubts so that others are not misled



### EXAMPLE

*Hey, Facebook friends! I wanted to address some information that has been circulating recently and share my doubts about its credibility. It's essential to be critical thinkers and not blindly believe everything we see or hear.*

*The reason I question the credibility of this information is due to the lack of important contextual details. The sources are unknown, and there's a missing context regarding the location, time, or any significant information that would help us better understand the situation. Such incomplete information makes it difficult to assess the accuracy and reliability of the claim.*

*The post seems to be only aimed at creating fear and fueling hate. Additionally, the lack of transparency regarding the perpetrators or hidden agendas raises further doubts about its authenticity.*

*As someone who values unbiased reporting, I encourage you all to be critical thinkers. It's essential to challenge our own views and look for facts, even if they challenge our preconceived notions.*

*I urge you all to actively participate in the comments section whenever you have doubts about the credibility of any information.*

*Remember, it's our collective responsibility to combat misinformation  
#CriticalThinking #SeekingTheTruth #FightingMisinformation  
#TogetherAgainstDisinformation*

---

---

---

---

---

---

---

---

---

---



### 3. Direct attack answer : follow these steps !



1. Be concise, factual and argumentative
2. Adopt a respectful tone, be firm and welcoming
3. Respond to criticism but not to personal attack
4. Acknowledge the emotion shared
5. Argument your answer with facts and sources and your intention or value
6. Thank them for engaging



#### Tone & style

Concise, factual, argumentative, respectful, firm, welcoming  
 "Thank you for your comments"

#### Respond

Respond to criticism but not to personal attacks  
 "In response to the concerns raised..."

#### Values

State your intention and values  
 "My intention was to highlight misinformation regardless of where I stand"

#### Listen

Acknowledge the emotion shared  
 "I understand that emotions can run high and we have differing perspectives"

#### Argument

Argument your answer with facts and sources  
 "Here are the facts and sources that support my stance"



## 4. Support someone who has been attacked: follow these steps



1. Counter singling a person out by depersonalizing the issue
2. Highlight similar views said by others
3. Reduce emotionality by zooming out to a big picture
4. Humanize the victim



### Depersonalize

Counter singling a person out by depersonalizing the issue

"Online attacks are not okay. I will take a stance regardless of whether I agree with the views expressed or not"

### Humanize

Humanize the victim

"There's a person behind the screen and our words carry weight"

### Regroup

Highlight similar views said by others

"This person's views represent a big group of people"

### Zoom out

Reduce emotionality by zooming out to a bigger picture

"There should be space for everyone to express themselves"



## 5. Make Memes with Memes Generator: super simple!



A meme is a humorous or thought-provoking image, video, piece of text, etc., that is copied, often with slight variations, and spread rapidly by internet users.

You can use online resources like <https://imgflip.com/memetemplates> to create memes in less than 1 minute.

1. Go to <https://imgflip.com/memetemplates>,
2. Pick an image you like, and click on add caption
3. Write your caption
4. Click on "Generate Meme" and take a screenshot or use the shared buttons provided.

### Examples to inspire you:

| Meme                                                                                                                                                                                                                                                  | Technics                                                                                         |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
|  <p data-bbox="581 1283 850 1440">The stadium was not polluted by Blue People</p> <p data-bbox="581 1566 850 1724">there is so much misinformation out there !</p> | <p data-bbox="1024 1276 1105 1304">1- Fact</p> <p data-bbox="1024 1644 1154 1671">2- Warning</p> |

**Meme**



**Technics**

Critical thinking questions presented in a fun way!

Easy to re-share.



Funny

Establishes new social norm

Highlights key, catchy questions that people can easily remember.

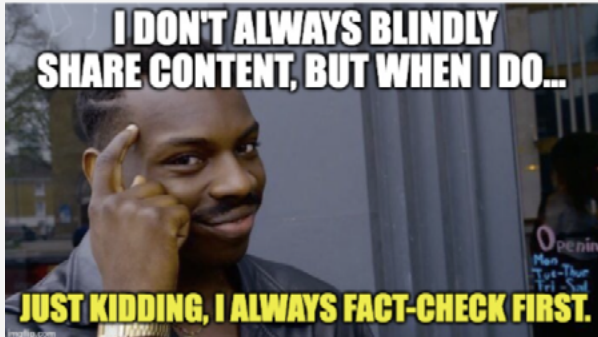


Funny

Establishes new social norm

Highlights key, catchy questions that people can easily remember.

## Meme



## Technics

Establishing new norms with humor.



Empathy prompt to establish new social norm, reduce propagation of hate speech and de-escalate with humour.





## 6. Take screenshots for Twitter when you don't have enough character space



Twitter has a character limit of 280 characters for each tweet, which can sometimes make it challenging to communicate complex ideas or share larger amounts of information.

When you have more to say than Twitter's character limit allows, one strategy you can use is to write your message in a note-taking app or text editor on your phone or computer, and then take a screenshot of that message.

Here's how you could do it:

1. Open a note-taking app on your device. This could be "Notes" on an iPhone, "Keep" on an Android, or any text editor on your computer.
2. Write out your message in full, making sure to check for clarity.
3. Once you're satisfied with your message, take a screenshot.
4. Now you can share the screenshot on Twitter as an image. Simply start a new tweet, attach the screenshot as a photo, and add any additional text you want to the tweet itself.



## 7. Use reverse image search on google



### An Easy Digital Tool: Reverse Image Search

#### What is Reverse Image Search and Why is it useful ?

Reverse image search is a search engine technology that uses an image as the query instead of text, enabling users to find where that image appears online, its origins, and other related information.

Reverse image search can be a powerful and very simple tool to fight mis information:

1. **Verification:** It helps in checking the source of an image or its occurrences elsewhere on the internet. This is particularly useful in debunking false news or misinformation, where a picture from a different event or context is presented with a new narrative.
2. **Contextual Clarity:** It can provide additional context to an image, revealing where it was first used and what discussion surrounded its initial deployment. This is helpful in uncovering misleading or deceptive uses of an image.
3. **Date and Location:** It often allows you to find out when and where the picture was taken, which can be crucial in fact-checking.
4. **Source Identification:** It can help identify the original creator or source of an image, which is useful for credit attribution or to find more credible information.
5. **Finding Similar Images:** It can show you visually similar images, which can assist in understanding the broader context or finding more relevant information.

#### A quick guide to reverse image search:



1. Go to a search engine such as Google, and click on the "Images" tab.
2. Click on the camera icon in the search bar to bring up the reverse image search tool.
3. You will have two options: either paste the URL of the image you want to search, or upload the image from your device.
4. Once you have uploaded the image or pasted the URL, click "Search".
5. The search engine will show you all the webpages where the image appears. This can help you identify the original source of the image or see if the image has been used elsewhere.

## 8. Use Checklists for yourself and share them with your friends



- Source Verification:** Have you verified the source of the information? Reliable information should come from trustworthy and reputable sources.
- Fact-checking:** Have you checked if the information is accurate and true? Use fact-checking websites to confirm the validity of the content.
- Emotional Check:** Does the post trigger strong emotions? Posts designed to incite anger, fear, or hatred may be forms of dangerous speech.
- Stereotype and Bias Check:** Does the content promote stereotypes or biases against certain groups or individuals?
- Harm Check:** Could the content potentially cause harm or distress to others? This could be physical, emotional, or psychological harm.
- Value Check:** Does sharing this content align with your values and the values of respect, peace, and understanding that you want to promote?
- Constructive Check:** Does the content contribute to a positive and constructive dialogue? If not, consider refraining from sharing it.
- Countercheck:** If the content is harmful but still needs to be shared (for awareness, for instance), are you providing counterspeech? Make sure to provide context, correct false information, and promote understanding in your counter-message. Be careful not to inadvertently repeat the myth when you are correcting it.

5

# Chapter 5



## Commit

I acknowledge that Dangerous Speech exists in our online and offline spaces, particularly in the forms of \_\_\_\_\_ (write the kind of Dangerous Speech you observe in your community: attack on minorities, disinformation campaign).

I believe these instances stem from \_\_\_\_\_ (write why you think this is happening: unchallenged biases, fear from war).

I am specifically committed to counteracting this form of speech by \_\_\_\_\_ (write how you want to take action: teach others, write messages).

To make this commitment measurable and time-bound, the first step I will take to address this is \_\_\_\_\_ (write a concrete action, such as posting educational content twice a week for the next three months to help my peers think before they share).

This action is achievable, given my resources and abilities, and it is relevant to promoting a more respectful and understanding community. I will track my progress by (method of measurement) \_\_\_\_\_."

6

# Chapter 6

## Be Safe out there



### Key message:

Ensuring your online safety is very important. It involves protecting personal data, managing digital footprints, respecting others' rights and privacy, and maintaining mental well-being. Navigating the internet safely means being aware of potential threats like misinformation and cyberbullying, while also harnessing the web's power responsibly and ethically.

### 1. Important resources to be safe online

1. Surveillance self defense guide by EFF: <https://ssd.eff.org/>
2. For common security scenarios the EFF has a set of tips and guides just for you. Check their scenario-based resources here: <https://ssd.eff.org/module-categories/security-scenarios>
3. In case of emergency, Access Now offers 24/7 support to activists in 9 languages (Arabic included) and responds within 2 hours through their hotline. <https://www.accessnow.org/help/>
4. Consumer Report's Security Planner is an excellent resource: <https://securityplanner.consumerreports.org/>
5. In case of emergency, look at this specific page: <https://securityplanner.consumerreports.org/tool/emergency-resources> it will point you to numerous hotlines and resources for people like you who might be experiencing online safety issues.
6. While data rates would apply to international texting, you can often find options to use whatsapp / online chats

## 2. Ten best practices from cyber security



|                                           |                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Protect your personal information</b>  | Avoid sharing sensitive personal information like your exact location, phone number, or financial details online. This can help protect you from potential threats or harassment.                                                                                                                                                                                                                               |
| <b>Be aware of your digital footprint</b> | Everything you share or post online can contribute to your digital footprint. Remember that once something is posted online, it can be difficult to completely remove it.                                                                                                                                                                                                                                       |
| <b>Don't engage with trolls</b>           | Trolls are individuals who purposely start arguments or post offensive comments to provoke and upset others. Engaging with them might lead to unnecessary conflict and feed into their harmful narratives.                                                                                                                                                                                                      |
| <b>Report and block abusive behaviors</b> | Use the reporting and blocking tools on social media platforms when you encounter abusive behavior or Dangerous Speech.                                                                                                                                                                                                                                                                                         |
| <b>Secure your accounts</b>               | Use strong, unique passwords for each of your online accounts. Consider using a password manager and enabling two-factor authentication whenever possible. Consider passwords managers, encryptions and apps like Signal for protecting your privacy.                                                                                                                                                           |
| <b>Inform trusted contacts</b>            | <p>Let someone you trust know if you're experiencing online harassment or dealing with Dangerous Speech. They can provide emotional support and help you report and document incidents.</p> <p>If you are experiencing serious threats or if it's simply too much, consider handing over your accounts to a 3rd person that you trust. Don't read all the dangerous messages to protect your mental health.</p> |
| <b>Check your privacy settings</b>        | Regularly review and update your privacy settings on different platforms to control who can see your posts and personal information.                                                                                                                                                                                                                                                                            |
| <b>Document everything</b>                | <p>If you are the target of Dangerous Speech and harassment, make sure you take screenshots, and document everything. This will be useful later for reporting the incidents to social media platforms, law enforcement or other organizations.</p> <p>Very important: if the attacks are serious and coordinated, you need to archive and document using the <a href="#">Wayback Machine</a>.</p>               |
| <b>Don't do it alone</b>                  | Activism is hard, don't do it alone. Lean on your community and work together with fellow activists. Solidarity and shared responsibility can often provide better security.                                                                                                                                                                                                                                    |
| <b>Call /escalate organization</b>        | If you're facing severe online abuse, you might need to escalate the issue to organizations that specialize in online safety and digital rights. They can provide guidance and support.                                                                                                                                                                                                                         |



### 3. P.E.A.C.E: Mental health for tackling Dangerous Speech



**Practice mindfulness and meditation.** Practices like meditation and deep-breathing exercises can be invaluable for managing stress and promoting mental well-being. They can help you stay centered and maintain a sense of balance amidst the chaos. Reflecting on your values can also help you manage your stress.

**Embrace boundaries.** Activism work can often feel like it's a 24/7 job. It's crucial to set boundaries around your time and availability. Designate "offline" hours for rest, relaxation, and disconnection from work and online platforms.

**Ask for help when needed.** If feelings of stress, anxiety, or depression are overwhelming, it's important to seek help from a mental health professional. Therapists and counselors can provide strategies to handle stress and prevent burnout. In addition, they offer a safe space to process your feelings and experiences. While these might not be easily available in challenging contexts or remote parts of the world, consider online resources such as the ones linked in this toolkit.

**Create a self-check routine.** Regular self-check-ins are crucial. Evaluate how you're feeling physically, emotionally, and mentally on a regular basis. If you're feeling overwhelmed, anxious, or perpetually exhausted, it may be a sign of burnout.

**Engage in activities that you enjoy.** Find the purpose and the joy in your work. No matter how challenging it is, a sense of purpose will help you.

