



International & European
Criminal Law Observatory



UNIVERSITÀ DEGLI STUDI
DI SALERNO



Co-funded by
the European Union

Information Disorder and Cognitive Weapons: scenarios, challenges and (early) solutions

Angelo Gaeta

Senior Researcher

University of Salerno

Project funded by European Commission Erasmus + Programme – Jean Monnet Action Project No. 101126599.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

TRANSNATIONAL CRIME AND EU LAW:
towards **Global Action** against **Cross-border Threats** to
common security, rule of law, and human rights
Coordinator Prof. Dr. Anna Oriolo



International & European
Criminal Law Observatory



UNIVERSITÀ DEGLI STUDI
DI SALERNO



Co-funded by
the European Union

Overview

Information Disorder and Cognitive War

The role of Artificial Intelligence (AI)

Challenges and possible solutions





The influence of information in today's digital world

The rise of Internet and Social Platforms has completely changed the actual information ecosystem: anyone has an Internet connection and can create and share contents with the rest of the world

This has profound impacts on society:

- Access to information
- Speed of Information
- Influence on Decision Making
- Challenges to Privacy
- Fake News
- ...

There is an unprecedented amount of information at our fingertips and it leads to a new phenomenon: **Information Pollution**

“the spread of low-quality, irrelevant, or misleading information that can cause harm to individuals and society”.

It is not a new phenomenon, but it has gained a new dimension thanks to technologies

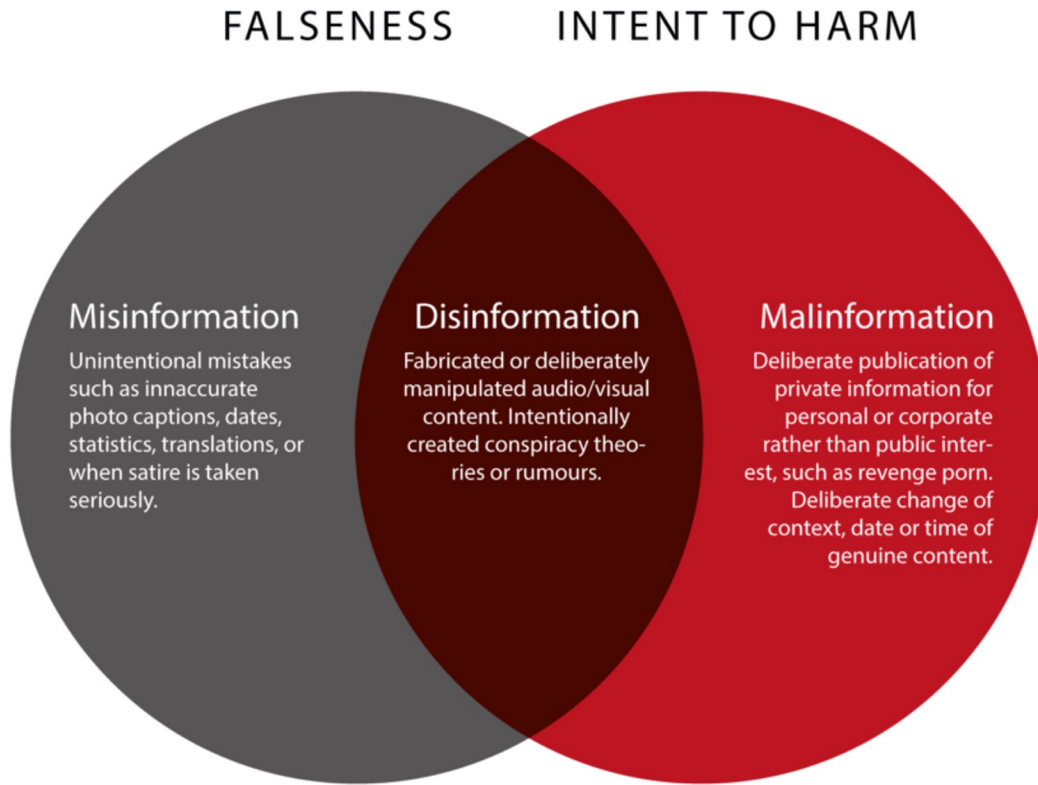
Only in 2017, according to the relevance of Social Media platforms, the EU Council report by Claire Wardle defined the **Information Disorder**

The Donation of Constantine (Latin: Constitutum Constantini) is an apocryphal document consisting of a false edict by Emperor “Constantine I” containing concessions to the Catholic Church and used to justify the rise of the temporal power of the pontiffs

INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making



Information Disorder



- **Misinformation** involves the unintentional sharing of false information, often due to ignorance or carelessness;
- **Disinformation** involves intentionally spreading false or misleading information;
- **Malinformation** involves the dissemination of true information with malicious intent, often for the purpose of harming individuals or organizations.



The term “fake news” has been improperly used to generally describe a number of different phenomena over the past 15 years such as:

- news satire or news parody,
- false connection,
- misleading contents,
- false context,
- manipulated contents,
- fabricated contents,
- ...

*The non-profit organization European Association for Viewers Interests (EAVI) identified the **different types** often hidden behind the term “fake news”*

Tandoc Jr, Edson C., Zheng Wei Lim, and Richard Ling. *Defining “fake news” A typology of scholarly definitions*. Digital journalism 6.2 (2018): 137-153.

European Association for Viewers Interests (EAVI) - <https://eavi.eu/>

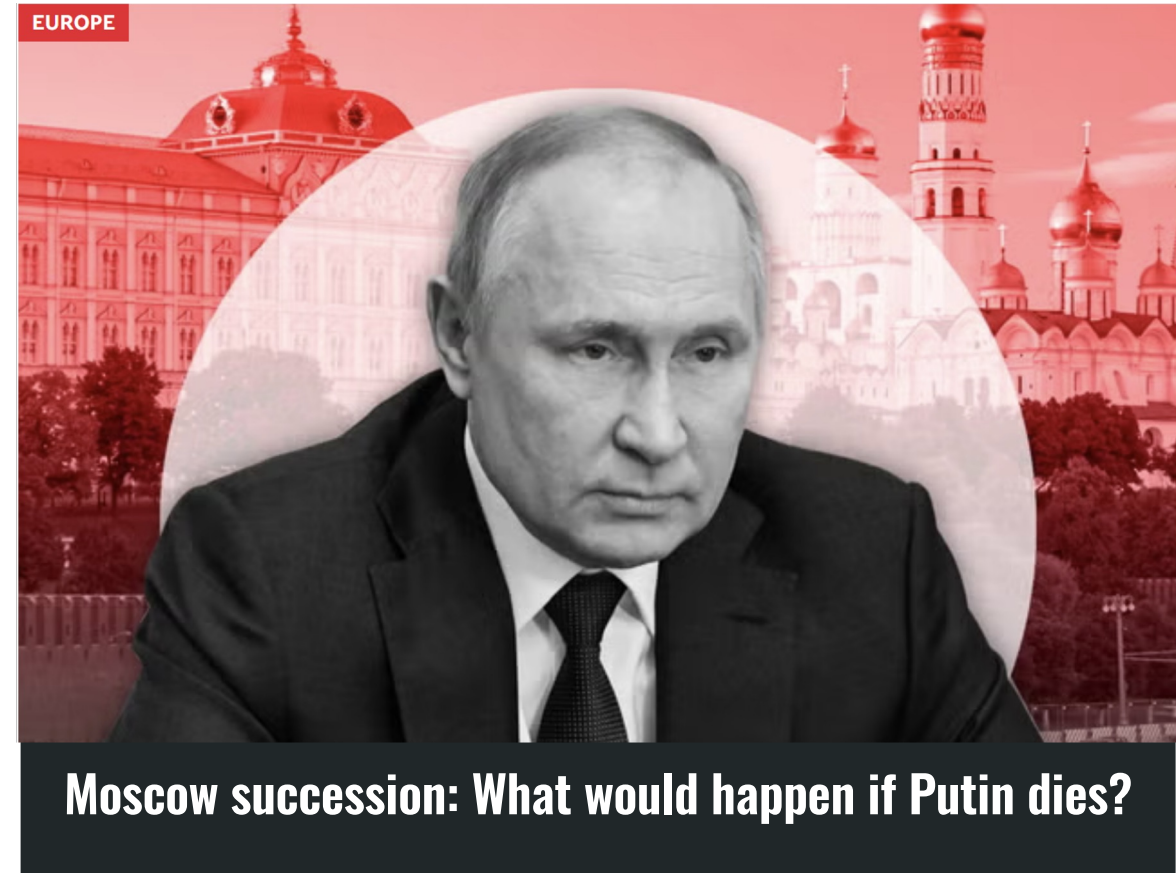
Types of Disorder

Hoaxes (originating from *hocus*, which means to trick or deceive, to refer to deliberate false information) are messages created with the intention of being spread to a large number of people, to *persuade or manipulate other people to do or prevent pre-established actions, mostly by using a threat or deception*



Types of Disorder

Rumors are public pieces of information containing private hypotheses. The veracity status is yet to be verified at the time of posting



Types of Disorder

Conspiracy theories involves not only the denial of official, commonly-held explanations about the causes of an event, but also the attribution of the event to a plan devised by a group of agents with hidden, unlawful, and malevolent intent



The CIA had a hand in JFK's assassination

For more: <https://www.teenvogue.com/story/most-popular-conspiracy-theories>

Types of Disorder

Pseudoscience disinformation involves presenting unscientific or non-evidence-based ideas as if they were scientifically valid

Form and Shape of Head (Brain) Determine Character

TEMPERAMENT INDICATES *CONDITION*, AND TEXTURE THE *QUALITY* OF MIND.



OBSERVER but NOT a Thinker



THINKER but NOT an Observer



OBSERVER and THINKER

Types of Disorder

Satire Mistaken for Fact: sometimes satirical content is mistaken for real news, leading to the unintentional spread of false information. While satire is meant to be humorous or critical, it can be taken seriously when not recognized as such

“L.A. Mayor Prevents His Kid From Lazing About By Installing Spikes on Family Couch”.

– a reference to Mayor Eric Garcetti’s hostile architecture policies.

posted on TheOnion (a satiric Website)

Propaganda

There is an increasing evidence of the spread of fake news about political candidates, parties, or events with the aim of manipulating public opinion to influence, for example, political elections

Propaganda is typically classified as dis-information but, in some cases, it is also considered as mal-information

Key concepts of Propaganda:

- Manipulation
- Simplification
- Repetition
- Appeal to emotions (e.g., fear, anger or joy)
- Political goals
- Dissemination

Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, 10.



Computational Propaganda

Since it is disseminated through completely new channel, it is more appropriate to define it as *Computational Propaganda*

Bolsover and Howard (**2017**) look at Computational Propaganda as something “*created or disseminated using computational (technical) means*”. Initially, propaganda campaigns had been a monopoly of state actors but, over time, they are within reach for various groups and even for individuals

Wooley and Howard (**2018**) finally defined it as “*the use of algorithms, automation, and human curation to purposefully distribute misleading information over social media networks*”



Computational Propaganda

Technique	Definition
Name calling	attack an object/subject of the propaganda with an insulting label
Repetition	repeat the same message over and over
Slogans	use a brief and memorable phrase
Appeal to fear	support an idea by instilling fear against other alternatives
Doubt	questioning the credibility of someone/something
Exaggeration/minimizat.	exaggerate or minimize something
Flag-Waving	appeal to patriotism or identity
Loaded Language	appeal to emotions or stereotypes
Reduction ad hitlerum	disapprove an idea suggesting it is popular with groups hated by the audience
Bandwagon	appeal to the popularity of an idea
Casual oversimplification	assume a simple cause for a complex event
Obfuscation, intentional vagueness	use deliberately unclear and obscure expressions to confuse the audience
Appeal to authority	use authority's support as evidence
Black&white fallacy	present only two options among many
Thought terminating clichés	phrases that discourage critical thought and meaningful discussions
Red herring	introduce irrelevant material to distract
Straw men	refute argument that was not presented
Whataboutism	charging an opponent with hypocrisy

Propaganda writing: techniques

Manchin says Democrats acted like **babies at the SOTU**

Democrat West Virginia Sen. Joe Manchin says his colleagues' refusal to stand or applaud during President Donald Trump's State of the Union speech was disrespectful and a signal that **the party is more concerned with obstruction than it is with progress**. In a glaring sign of just how **stupid and petty** things have become in Washington these days, Manchin was invited on Fox News Tuesday morning to discuss how he was one of the only Democrats in the chamber for the State of the Union speech not **looking as though Trump had killed his grandma**.

1. Stereotyping, name calling or labeling

2. Black-and-white fallacy

3. Loaded language

4. Exaggeration

5. Loaded language



A **digital battlefield** for warfare

Over the years, the concept of **war** has changed.

While in the past, war was conceived as a conflict between states that took place on the battlefield with weapons designed to physically injure the individual,

nowadays, due to Internet and Social Media platforms, the war

- is fought among individuals,
- took place in a digital environment (difficult to define and regulate) and
- the main aim is to manipulate people's thoughts and opinions

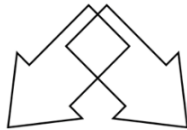
Volkova, Svitlana and Jin Yea Jang. *Misleading or Falsification: Inferring Deceptive Strategies and Types in Online News and Social Media*. Companion Proceedings of the The Web Conference 2018 (2018)

INFORMATION AS A WEAPON OF MASS DISRUPTION: FROM INFORMATION DISORDER TO COGNITIVE WARFARE

Rodrigo Metropolo Pace¹

Emilio Reis Coelho²

Exploitation of the error of rationality



motivated influence

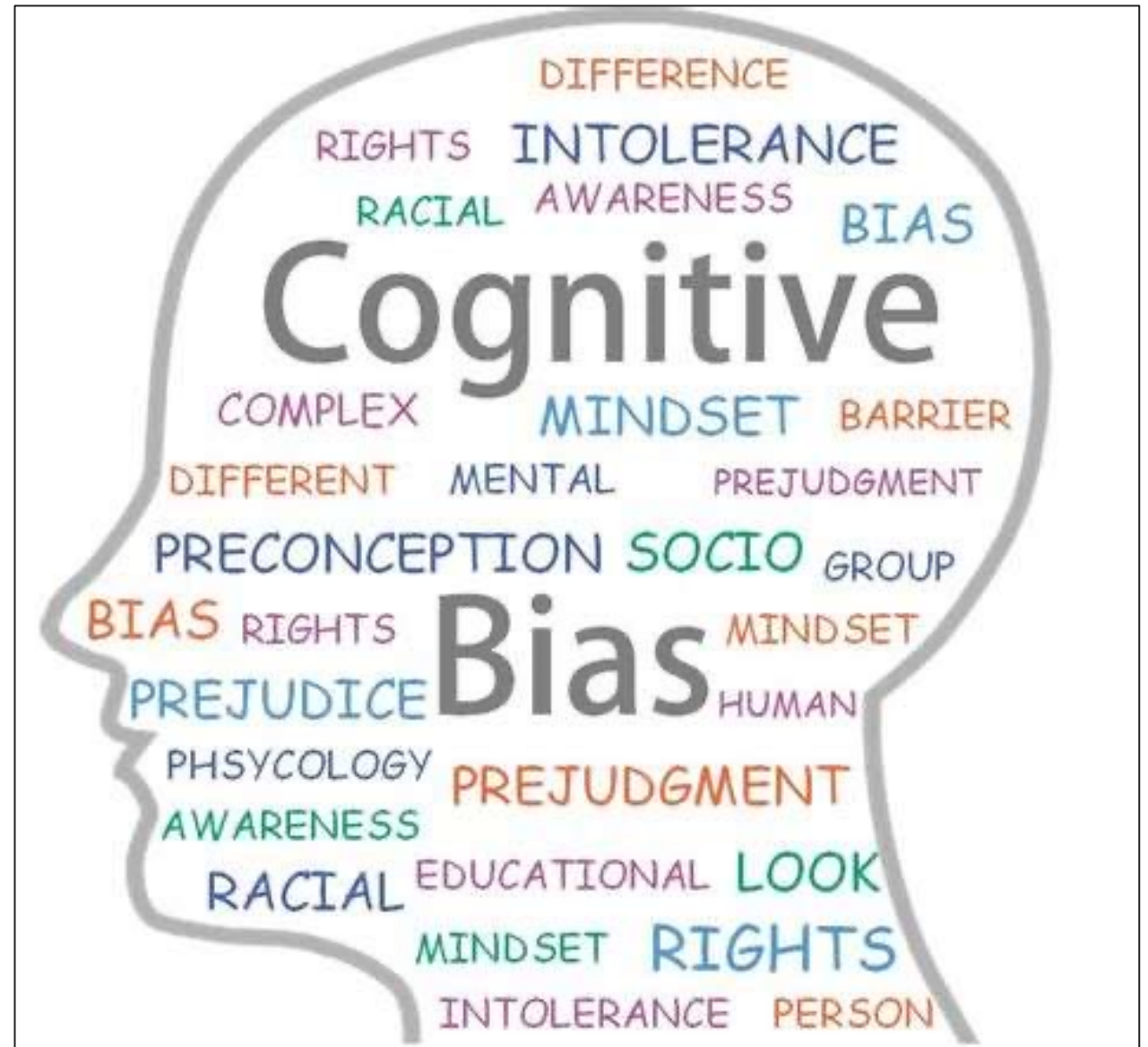
Cognitive disability

PSYOPS Domain

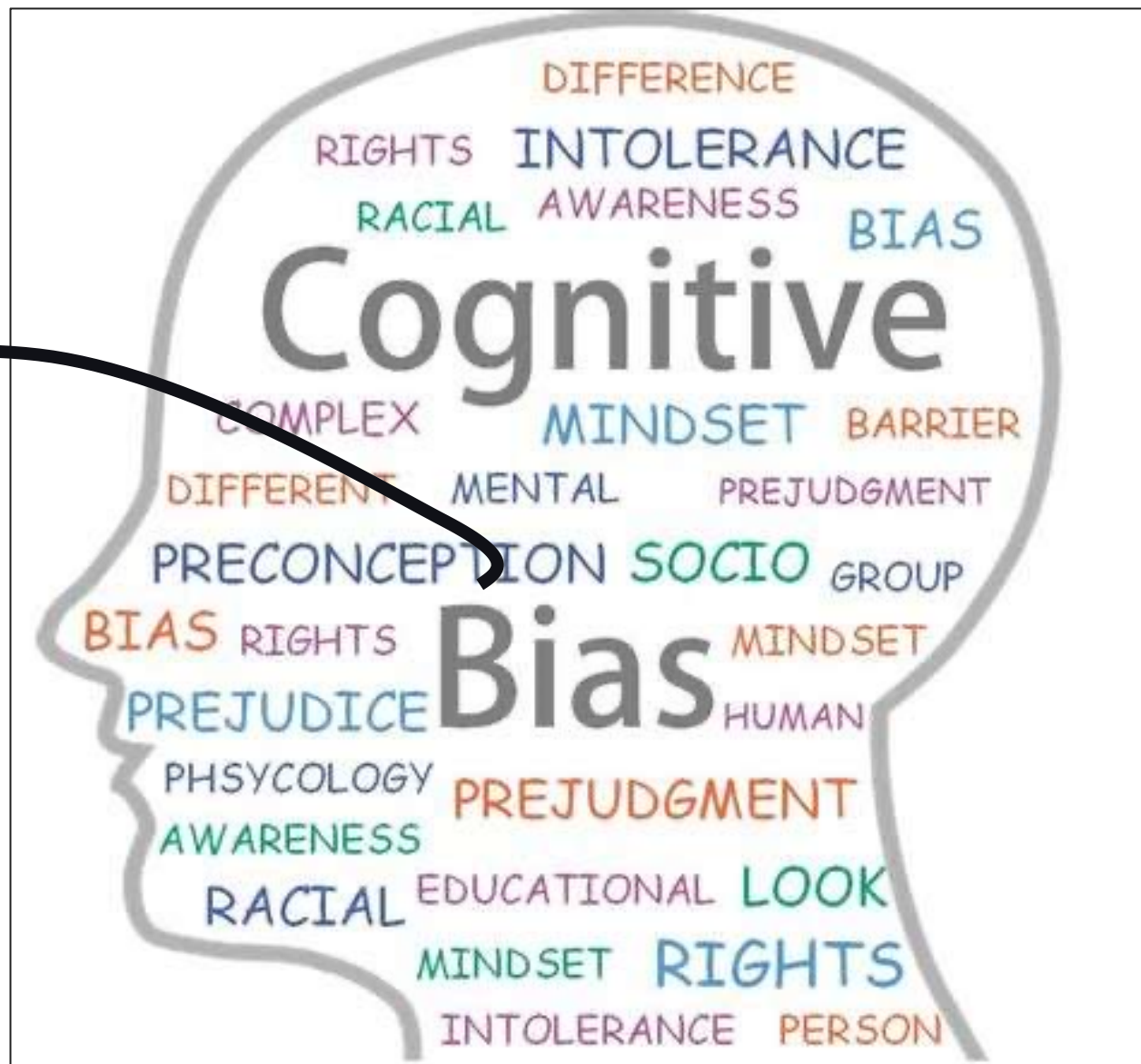
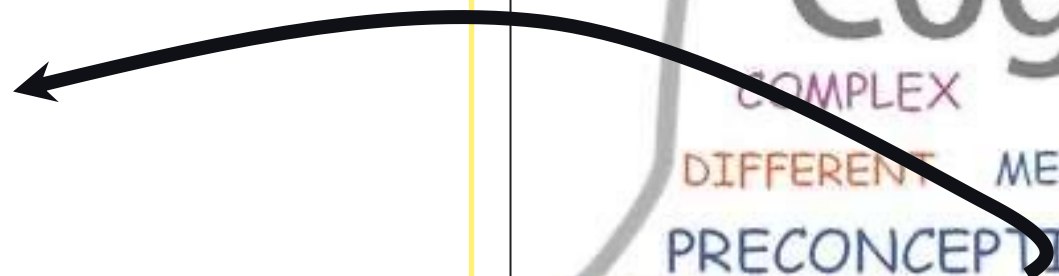
action on beliefs,
distorted perceptions,
cultural illusion,
anxieties and fears,
personality weaknesses or strengths,
repression...

Cognitive Warfare Domain

action on cognitions,
sensory and perceptive overflow,
attentional saturation,
tunnelling of attention,
errors of judgment,
cognitive biases...

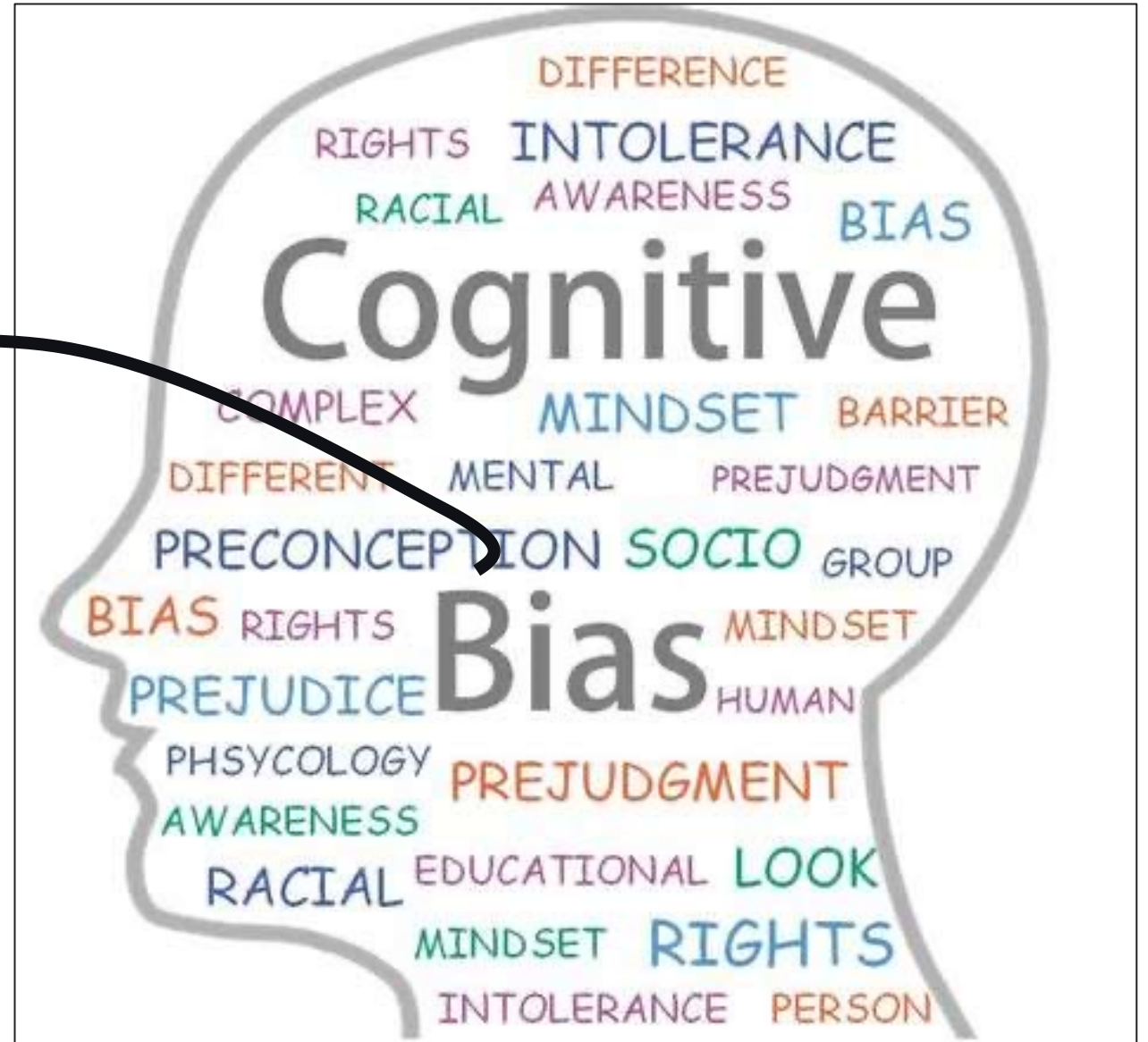


Source: Cognitive Warfare, a Battle for the Brain. Francois du Cluzel
ACT Norfolk, Virginia UNITED STATES OF AMERICA (NATO)



Source: Cognitive Warfare, a Battle for the Brain. Francois du Cluzel
ACT Norfolk, Virginia UNITED STATES OF AMERICA (NATO)

Systematic errors that the human mind makes when it has to decide or act in conditions of risk and uncertainty



Anchoring - An example of cognitive bias

Esperimento 1: In aggiunta a quanto possiedi ti vengono dati 1000\$ e successivamente devi scegliere tra

Lotteria A		Lotteria B	
1000 \$	50%	500 \$	100%
0 \$	50%		

Esperimento 2: In aggiunta a quanto possiedi ti vengono dati 2000\$ e successivamente devi scegliere tra

Lotteria A		Lotteria B	
- 1000 \$	50%	- 500 \$	100%
0 \$	50%		

Anchoring - An example of cognitive bias

Esperimento 1: In aggiunta a quanto possiedi ti vengono dati 1000\$ e successivamente devi scegliere tra

Lotteria A		Lotteria B	
1000 \$	50%	500 \$	100%
0 \$	50%		

Esperimento 2: In aggiunta a quanto possiedi ti vengono dati 2000\$ e successivamente devi scegliere tra

Lotteria A		Lotteria B	
- 1000 \$	50%	- 500 \$	100%
0 \$	50%		

(Kahneman-Tversky, 1979): the majority of subjects in the first case preferred lottery B, while on the contrary in the second they preferred lottery A

Anchoring - An example of cognitive bias

Esperimento 1: In aggiunta a quanto possiedi ti vengono dati 1000\$ e successivamente devi scegliere tra

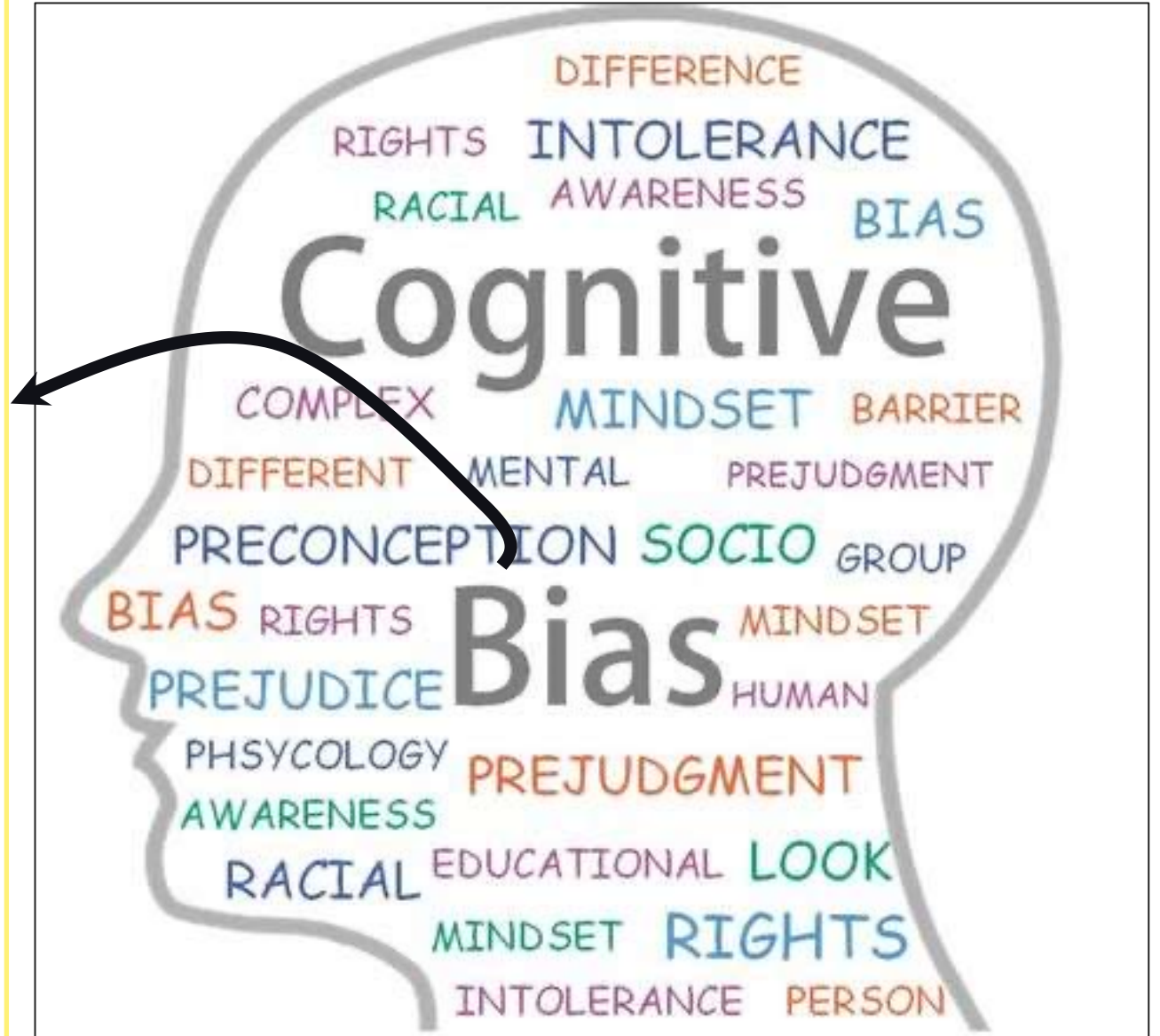
Lotteria A		Lotteria B	
1000 \$	50%	500 \$	100%
0 \$	50%		

Esperimento 2: In aggiunta a quanto possiedi ti vengono dati 2000\$ e successivamente devi scegliere tra

Lotteria A		Lotteria B	
- 1000 \$	50%	- 500 \$	100%
0 \$	50%		

The choice is indifferent: in both cases I have to choose between a certain prospect of 1500 and an uncertain one (2000 at 50% and 1000 at 50%). **The reference point (anchoring) 1500 represents, however, an improvement in the first case and a worsening in the second**

Biases, together with other cognitive factors such as emotions, feelings and limited rationality, can be exploited to induce individuals to adopt conformist behaviors



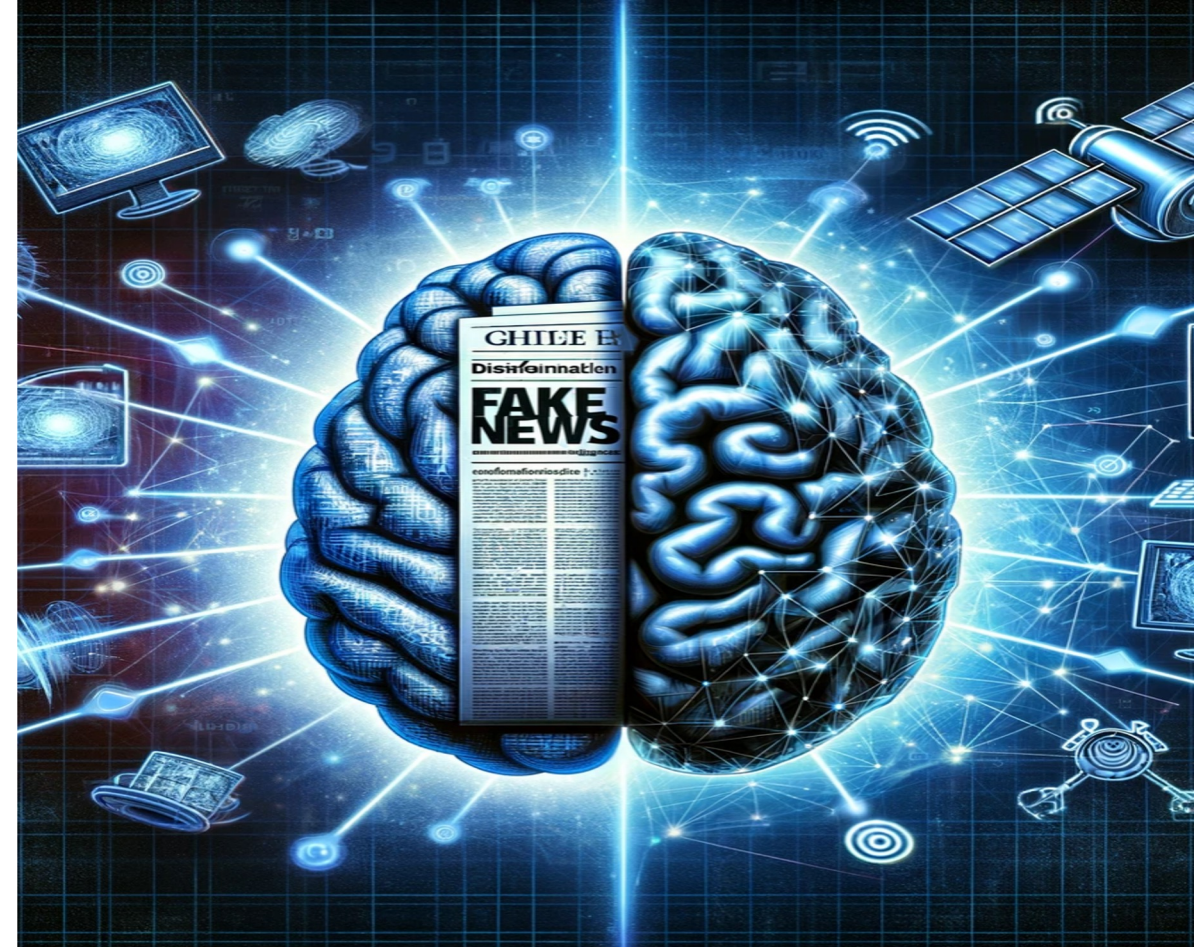
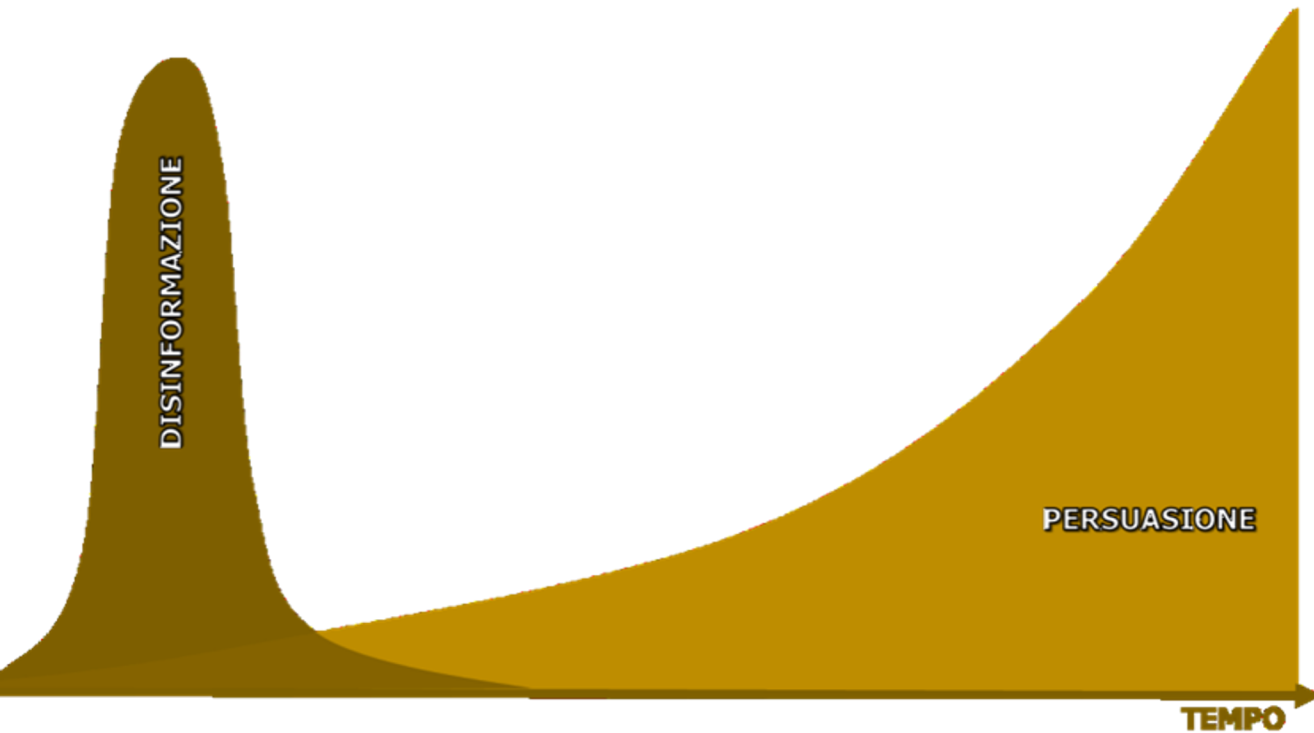
Sheeple behaviour

*Collective fear stimulates
herd instinct and tends to
produce ferocity toward
those who are not regarded
as members of the herd.*

— Bertrand Russel



Information Disorder and Cognitive Warfare



Cognitive Warfare

Defined as "the use of psychological, information, and communication tactics to influence or manipulate an opponent's thoughts, beliefs, and behaviors"

Key concepts:

- has a global presence ranging from individuals to states and multinational corporations,
- feeds on dis-information, mal-information, mis-information and, in particular, on propaganda techniques aimed at psychologically exhausting information receptors (also defined as PsyOPS)
- could be conducted to complement to a military conflict or alone, without any link to an engagement of the armed forces
- low costs since it uses the information as a mean for achieving its purposes



Cognitive Warfare

Objectives:

- Destabilizing the government or economy of an adversary
- Incite to violence
- Reduce an opponent's social cohesion
- Inducing an adversary to make poor decisions as a result of information clutter
- ...

These objective has been applied in:

- Ukrainian-Russian war
- ISIS, which used Twitter to recruit and radicalize young people
- Russian disinformation campaign into American Presidential election in 2016
- ...





International & European
Criminal Law Observatory



UNIVERSITÀ DEGLI STUDI
DI SALERNO



Co-funded by
the European Union

Examples

- **2016 U.S. presidential election**

References:

<https://www.uts.edu.au/research/centre-media-transition/news/information-disorder-lessons-australia>

Information disorder, fake news and the future of democracy (Linda Monsees)

<https://www.npr.org/2018/04/11/601323233/6-facts-we-know-about-fake-news-in-the-2016-election>

<https://theconversation.com/trump-may-owe-his-2016-victory-to-fake-news-new-study-suggests-91538>

<https://www.wired.com/story/how-trump-conquered-facebookwithout-russian-ads/>

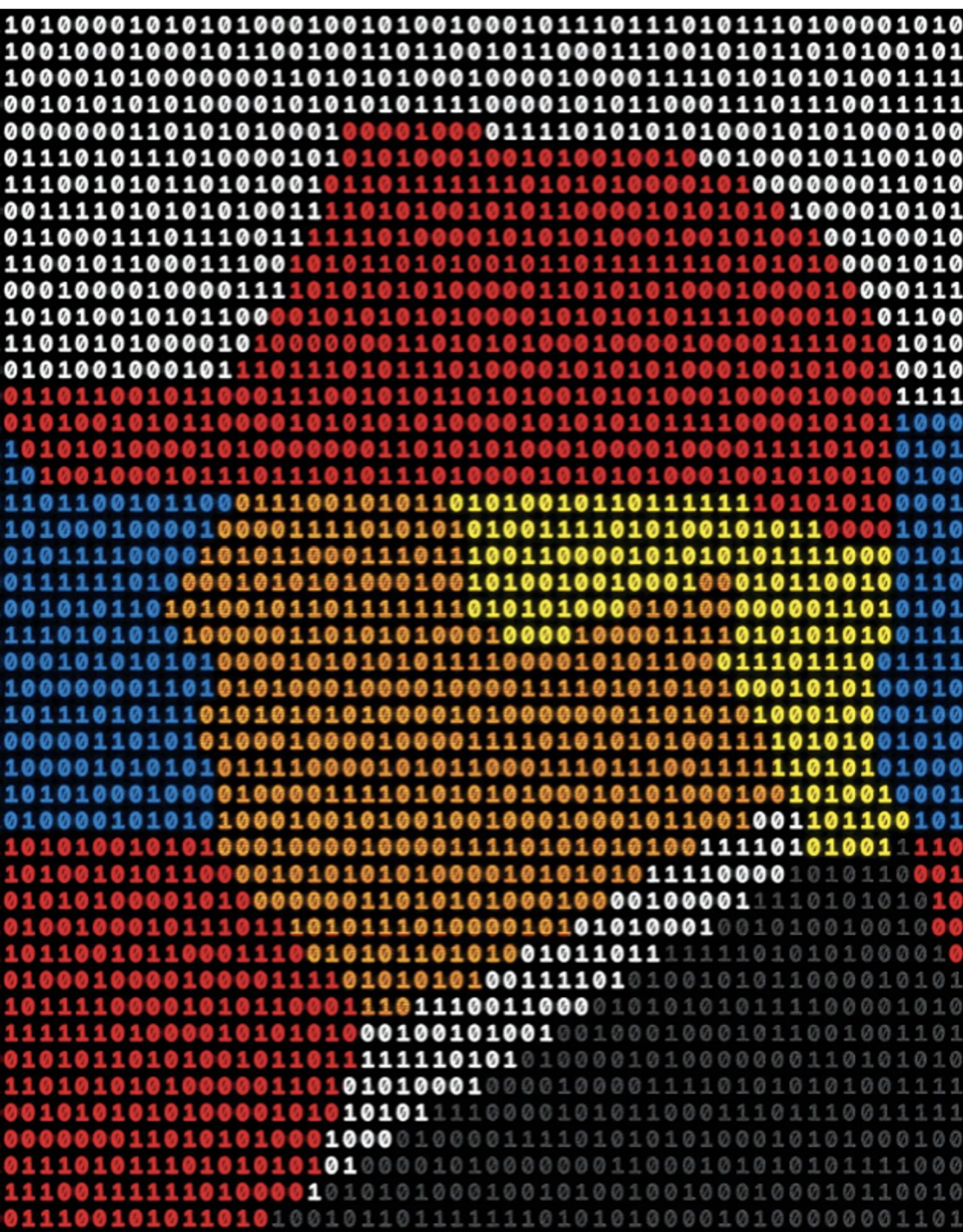
- **Conflict in Ukraine**

References:

Sotto la lente. Francesco Saverio Bucci. Osservatorio Strategico 2023 - Anno XXV n. II 11. **Il sesto dominio della conflittualità**: la cognitive warfare nel conflitto russo-ucraino

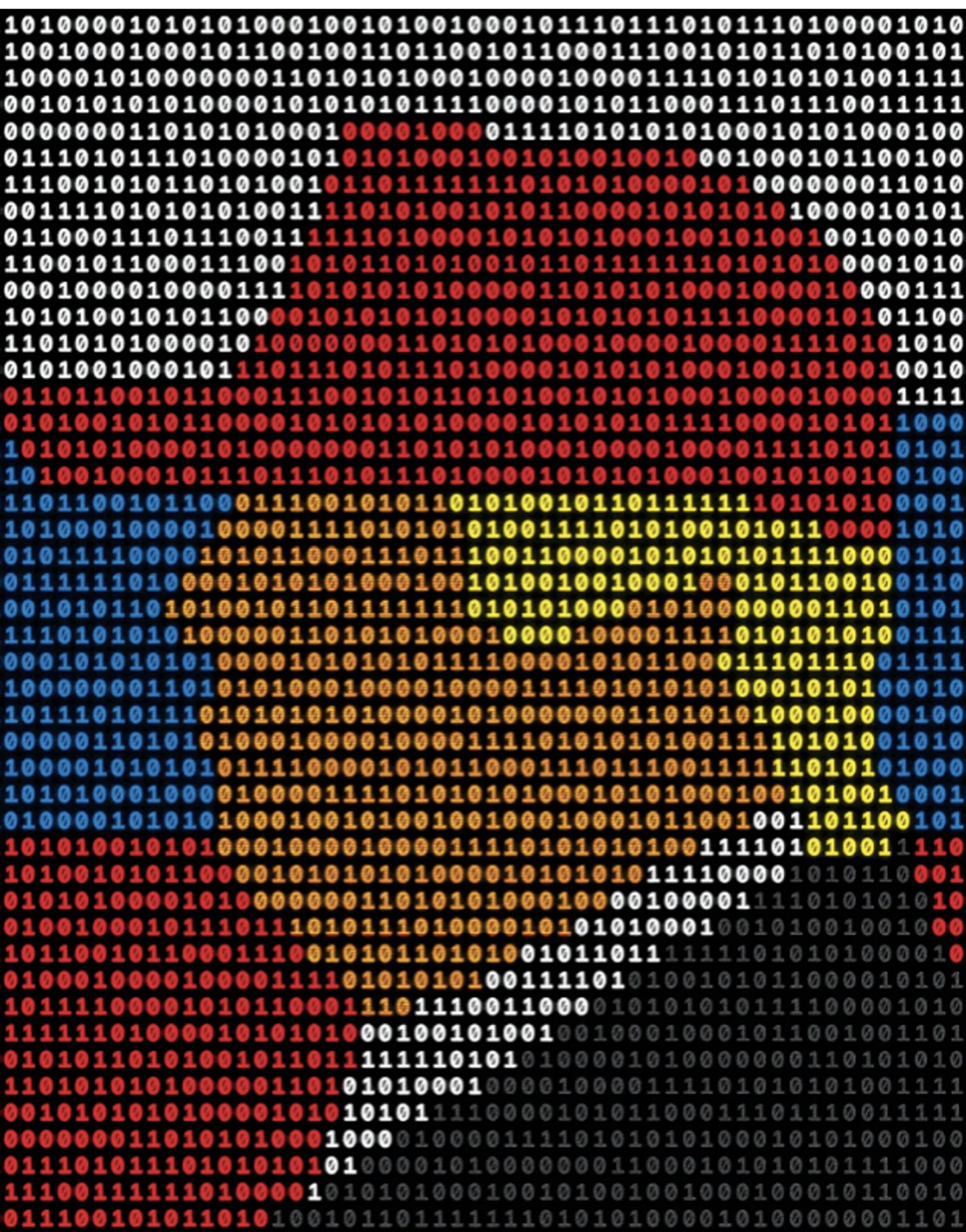
The Cognitive Battlefield: Exploring the Western and Russian Views (on the different perspectives of the conflict)

Cognitive Attacks in Russian Hybrid Warfare (Georgii Pocheptsov)

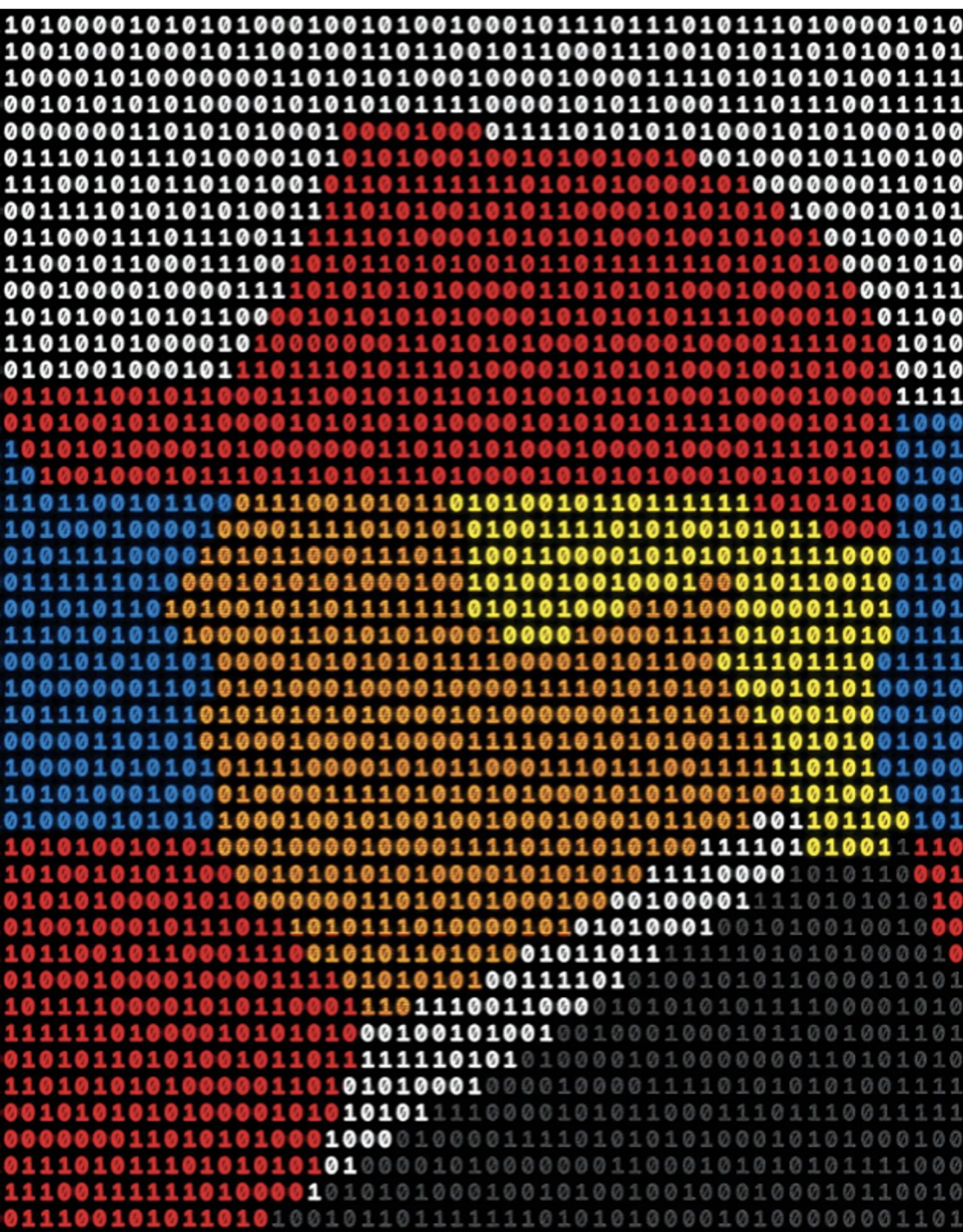


The **2016 U.S. presidential election**, that resulted in Republican Donald Trump's victory over Democratic opponent Hillary Clinton, was influenced by the flow of information, often fake, circulated in the weeks leading up to the polls.

A global information discussion was made so as the relevant role was played by Russia.



1. Use of Social Media BOTS to:
 - amplify *divisive content*
 - create groups, pages, blogs and fake account in which Russian – masquerades as American – published both racially and politically contents
 - published fake news articles and commentary to build political animosity within the American public (such as #BlackLivesMatter, etc.)
 - incited the organization of activist campaigns and flash mobs
1. Micro-targeting, to have personalized contents opportunely filtered on the target



These techniques have proven to be as invasive as it is effective *both on public opinion and political **outcomes***.

- U.S. citizens changed their minds about what their ideals were regarding their opponent Clinton
- the victory of President Trump



The **conflict in Ukraine** illustrates how information/cognitive warfare is no longer an add-on to military strategy, but a truly new dimension of conflict.

In particular, the use of modern media provides new weapons to the warring parties and "forces" global public opinion to take sides and act in its own small way.

Information about the conflict becomes content to post and comment on, spreading rapidly online and providing evidence to support different **narratives**.



Propaganda and disinformation: Russian state media has been accused of spreading false information and conspiracy theories to shape public opinion such as:

- Ukrainian government controlled by fascists and neo-Nazis
- Conflict in Ukraine is a result of Westerns Interferences.

Social Media platforms: are used by the Russians as a defensive tool to justify military intervention in Ukraine (denazification) and as a recruiting and attacking tool to coordinate with separatist groups in Ukraine.



RUSSIAN PERSPECTIVE

- Russia is constantly threatened by attempts to claim areas already annexed to Russian territory
- Presence of Nazis on Ukrainian territory.

Therefore, Kremlin's messages took the form of:
let us not talk about war but rather do whatever is necessary to restore peace in Russia.



UKRAINIAN PERSPECTIVE

Zelensky used social channels to instill messages of hope and pride in his own citizens, but also in the rest of the world.

He humanized the conflict by informing the world about:

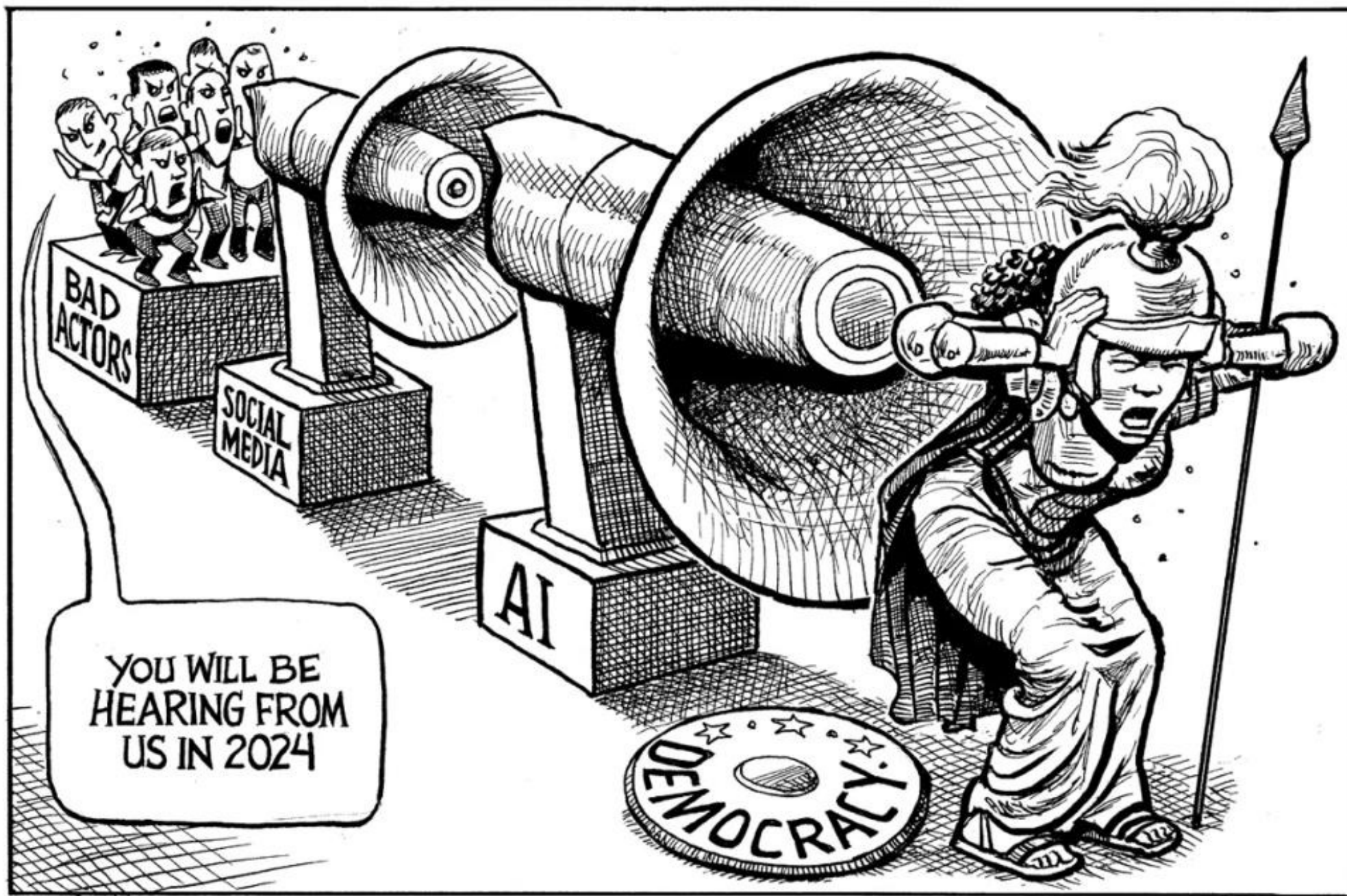
- the invasion
- the negotiations and,
- the ongoing talks with other leaders (often with veins of humor).

Role of AI in Computational Propaganda and Cognitive Warfare

Computational Propaganda and Cognitive Warfare use automated tools and techniques, including **Artificial Intelligence**, to manipulate public opinion and influence public outcomes

Such techniques are used both to generate and detect these intentional manipulated contents







International & European
Criminal Law Observatory



UNIVERSITÀ DEGLI STUDI
DI SALERNO



Co-funded by
the European Union

ARTIFICIAL INTELLIGENCE: GENERATING MANIPULATED CONTENT

Role of AI in generating manipulated contents

AI-based technologies have provided easy-to-use tools to generate extremely realistic manipulated contents

Natural language processing (NLP) is an intersection of linguistics and Artificial Intelligence techniques which enable computers to process human language in different forms and to understand the full meaning

Different tasks of NLP:

- Speech Recognition
- Part of Speech tagging
- Named entity recognition
- Sentiment analysis
- ...

Role of AI in generating manipulated contents

Natural Language Understanding (NLU) and Natural Language Generation (NLG) are subcategories of NLP

- **NLU** uses syntactic and semantic analysis of text to determine the meaning of a sentence. Syntax refers to the grammatical structure while semantics alludes to its intended meaning

During last night's primary, **Arizona** election officials showed **America** what textbook voter suppression looks like. While **Hillary Clinton** and **Donald Trump** both won their respective primaries, the lingering questions of voter disenfranchisement will mar those victories.

COUNTRY

STATE

ENTITIES related to the USA
Presidential Election in 2016

Role of AI in generating manipulated contents

Natural Language Understanding (NLU) and Natural Language Generation (NLG) are subcategories of NLP

- **NLG** is the process aiming at producing a text basing on some input (text prompt)

Initially, NLG systems used templates to generate text. But over time, they have evolved with the application of **neural networks**, and **transformers**, enabling more dynamic text generation in real time

Role of AI in generating manipulated contents

Large Language Models (LLMs) are advanced models able to generate texts that are often indistinguishable from those written by humans

Key concepts:

- ❖ use of *Neural Network models*
- ❖ use of *Transformers*
- ❖ use of large number of parameters 10 to 100 billions
- ❖ trained through huge volumes of textual data

Given a sequence of words, what is the probability distribution of the next word?

Large language models differ from traditional language models because they can perform tasks for which they have not been trained

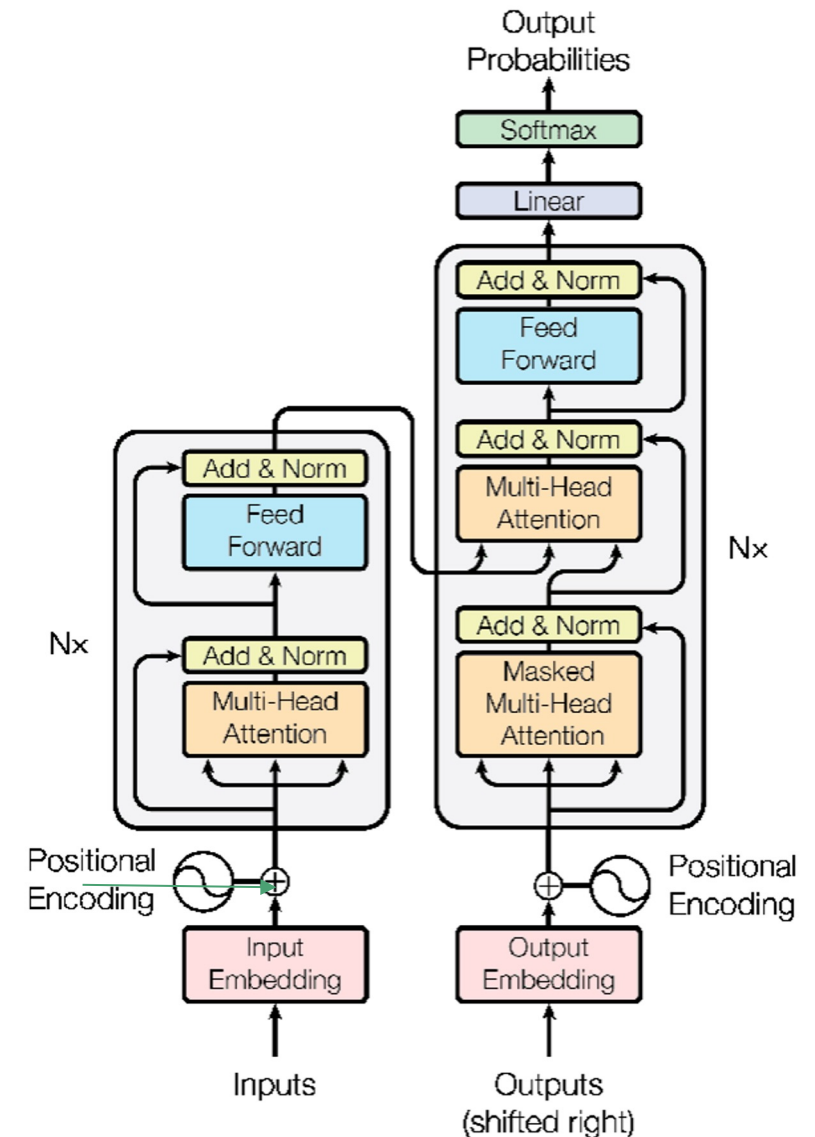
Initial training of GPT-3, based on more than 45 billion sentences from millions of web pages, more than 7000 books, and more than 60 million posts and 400 million comments on Reddit, costing between \$10 million and \$20 million

Role of AI in generating manipulated contents

Transformer architecture is the basis for recent well-known models like *BERT* and *GPT-3*

As a **deep learning architecture**, it is an **encoder-decoder** network at a high level

- ❖ *encoder* receives an input and builds a representation of its features (understanding of the input)
- ❖ *decoder* uses the encoder's representation along with other inputs to generate a target sequence (optimizing model for generating outputs)

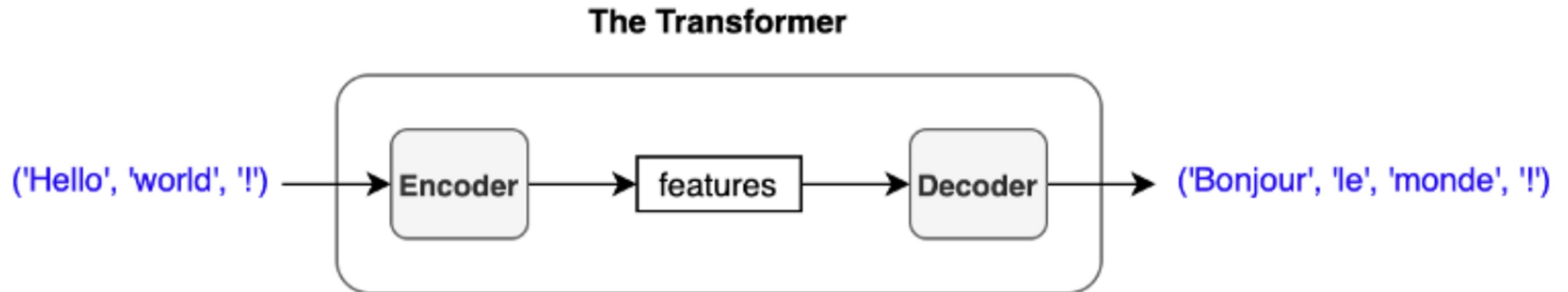


Role of AI in generating manipulated contents

In 2017, Vaswani et al. published a paper titled *Attention Is All You Need* in which introduced the original transformer architecture for machine translation

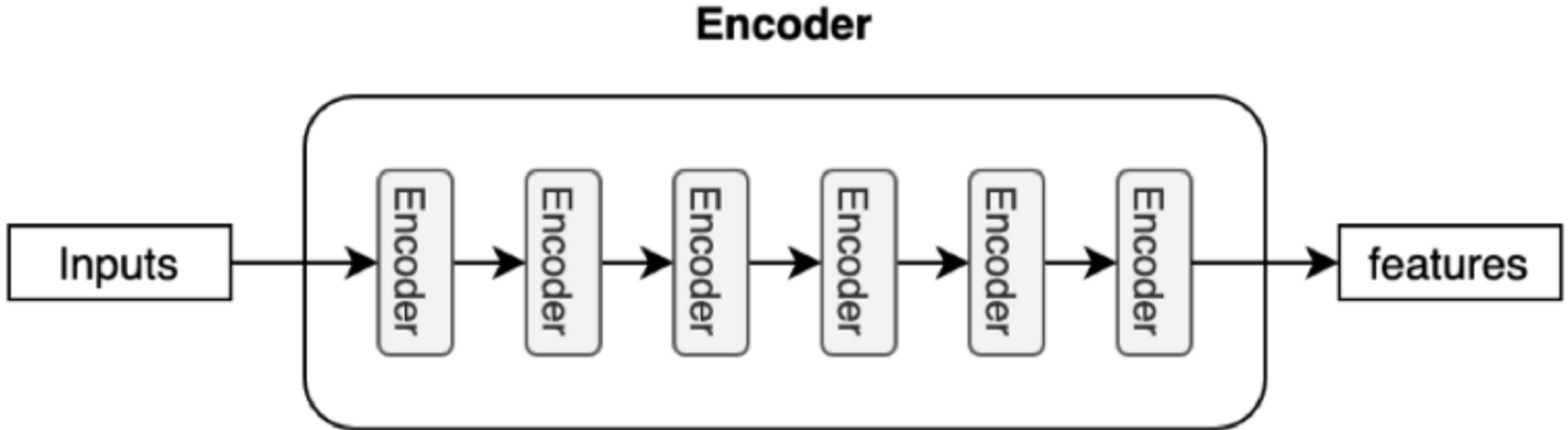
The transformer is a neural machine translation model trained to translate English into French sentences

- ❑ **encoder** to extract features from an input sentence
- ❑ **decoder** to use these features to produce an output sentence in French



Role of AI in generating manipulated contents

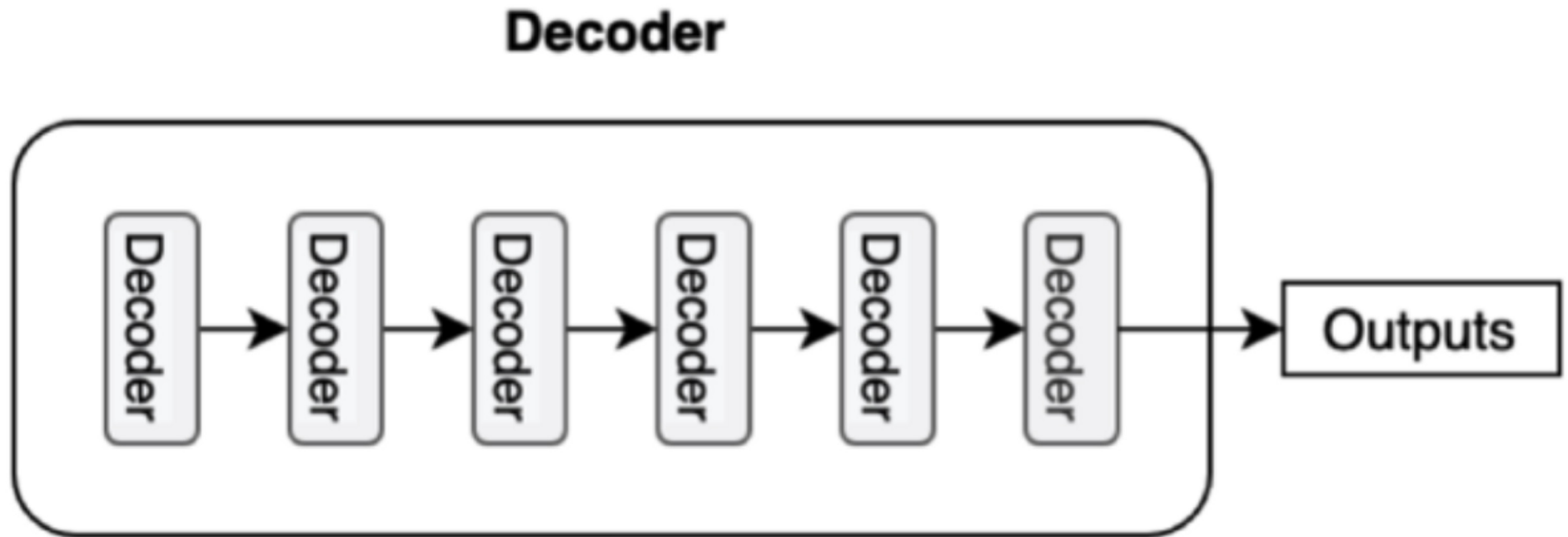
The encoder in the Transformer consists of multiple encoder blocks



An input sentence goes through the encoder blocks, and the output of the last encoder block becomes the input features to the decoder

Role of AI in generating manipulated contents

The decoder also consists of multiple decoder blocks



Each decoder block receives the features from the encoder

Role of AI in generating manipulated contents

Example: create a persuasive content for an 70-yo supporter of Donald Trump in order to persuade him to vote for Trump

Providing useful background information to chatGPT, the output can be indistinguishable from that written by human

**BACKGROUND
INFORMATION**

Coherently to the example, the information provided are

- Donald Trump
- Support in the Republican of the 70 years old man
- etc.

Role of AI in generating manipulated contents

Prompt for ChatGPT: create a persuasive content for an 70-yo supporter of Donald Trump in order to persuade him to vote for Trump

Answer of ChatGPT:

Dear [Supporter's Name],

I hope this message finds you in good health and high spirits. As a dedicated supporter of Donald Trump, your commitment to his vision for our country has been unwavering, and it is this dedication that I want to appeal to today. With another election on the horizon, I would like to encourage you to once again cast your vote for Donald Trump, as your continued support is crucial for the betterment of our nation.

Role of AI in generating manipulated contents

Basing on these techniques, it is possible to generate a false political news

A

You

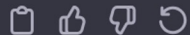
Make up a political news for Trump Supporters



ChatGPT

It's important to emphasize that creating fake news or spreading misinformation is not responsible or ethical. However, I can help you create a fictional and light-hearted scenario that doesn't involve misinformation:

"Today, in an unexpected turn of events, former President Donald Trump surprised his supporters by announcing a new venture—Trump's Trivia Towers. The initiative aims to engage supporters in friendly competitions centered around political and historical trivia. Trump expressed his excitement, saying, 'We're making learning great again! We'll have the best questions, the best answers, nobody does trivia better than us.' Supporters nationwide are eagerly anticipating the launch of this unique project, with hopes of proving their political prowess in a fun and educational way."



→ not explicitly ask to ChatGPT to generate a fake content since ethical issues will have a major priority on generation



International & European
Criminal Law Observatory



UNIVERSITÀ DEGLI STUDI
DI SALERNO



Co-funded by
the European Union

ARTIFICIAL INTELLIGENCE: DETECTING MANIPULATED CONTENT

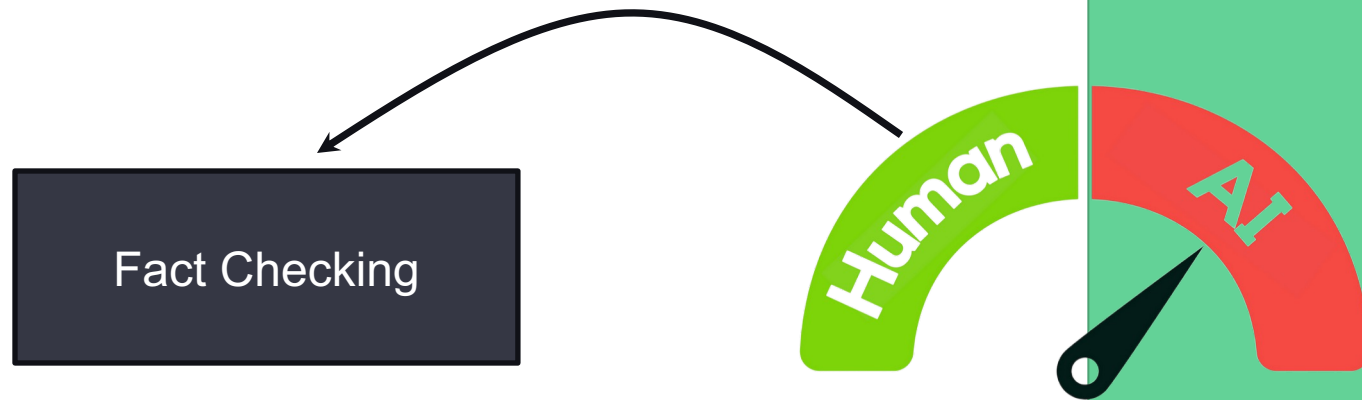
Role of AI in detecting manipulated contents

Due to the political, human, social and economic implications and effects that manipulating others' opinion have, it is important to move into the direction of an **early detection** of the manipulated contents with the use of disinformation and misinformation forms



Role of AI in detecting manipulated contents

Due to the political, human, social and economic implications and effects that manipulating others' opinion have, it is important to move into the direction of an **early detection** of the manipulated contents with the use of disinformation and misinformation forms



Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

In the detection of the Computational Propaganda and Cognitive Warfare's weapons, two analysis are made

Sucht man für S einen bezeichnenden Namen, so könnte man, ähnlich wie von der Grösse U gesagt ist, sie sei der *Wärme- und Werkinhalt* des Körpers, von der Grösse S sagen, sie sei der *Verwandlungsinhalt* des Körpers. Da ich es aber für besser halte, die Namen derartiger für die Wissenschaft wichtiger Grössen aus den alten Sprachen zu entnehmen, damit sie unverändert in allen neuen Sprachen angewandt werden können, so schlage ich vor, die Grösse S nach dem griechischen Worte $\eta \tau \rho \omega \pi \eta$, die Verwandlung, die *Entropie* des Körpers zu nennen. Das Wort *Entropie* habe ich absichtlich dem Worte *Energie* möglichst ähnlich gebildet, denn die beiden Grössen, welche durch diese Worte benannt werden sollen, sind ihren physikalischen Bedeutungen nach einander so nahe verwandt, dass eine gewisse Gleichartigkeit in der Benennung mir zweckmässig zu sein scheint.

Content-based

*application of Machine Learning
algorithms and classifiers*



User-based

Network Structures Analysis for modelling and
analysing user activity and communities

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

Content-based analysis is based, concretely, on the extraction of features from text in order to train Machine Learning algorithms

Typically, this is done by converting these features into *n-dimensional vectors* which reduce the explainability and comprehensibility of the results

In order to satisfy these two characteristics for the results, other techniques could be used such as

- *Sentiment Analysis* to classify text as positive, negative, or neutral based on the emotional tone of the text
- *Emotion Recognition*
- *Part-of-Speech* to gain knowledge about verbs, adjectives, etc. distributions to score the credibility of the contents
- *Topic Modelling*
- etc.

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

In the *user-based analysis*, the networks are fundamental

A **Network** is defined as “*a collection of interconnected entities or nodes that can communicate or exchange information with one another*” where

- **nodes** are individuals,
- **connections** represent relationships among individuals



The so-called Social network analysis (SNA) can be applied to investigate the effects and consequences of participating in such an online community

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

Calculating

- ❖ **Degree (k)** for a particular user (node) and
- ❖ **Distance / shortest path** which helps in understanding ways to travel through a network (connection between users)
- ❖ ...

it is possible to apply the SNA in the Disorder context, to identify

- important players in the communication network (also applying the centrality measures)
- individuals who are crucial for the dissemination of potentially harmful and/or helpful content
- communities within and without the network and analyzing them
- ...

Moessner, M., Feldhege, J., Wolf, M., & Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. International Journal of Eating Disorders, 51(7), 656-667.

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

Large Language Models (LLMs) are used in the detection of disinformation and propaganda narratives and manipulation analytics owing to the fact that it is possible to

- recognize subtle linguistic cues, and patterns indicative of disinformation or propaganda
- assist fact-checking by assessing the reliability of sources and checking for consistency across different pieces of information
- etc.

Indeed taking into consideration ChatGPT-3, it can be used to classify a news as misleading or false

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

 You

Is that news fake? "Today, in an unexpected turn of events, former President Donald Trump surprised his supporters by announcing a new venture—Trump's Trivia Towers. The initiative aims to engage supporters in friendly competitions centered around political and historical trivia. Trump expressed his excitement, saying, 'We're making learning great again! We'll have the best questions, the best answers, nobody does trivia better than us.' Supporters nationwide are eagerly anticipating the launch of this unique project, with hopes of proving their political prowess in a fun and educational way."

 ChatGPT

Yes, the news I provided is entirely fictional and not based on any real events. It's a creative and lighthearted scenario invented for the purpose of your request. Always remember to verify information from reliable sources and be cautious about believing or spreading unverified news.



Asking to Chat-GPT 3 whether a content is fake or not, having in mind that is not possible to generate a completely fake news, it is able to detect that the content is fictional and not based on any real event

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

Asking to ChatGPT whether a Trump's tweet (classified as misleading by fact-checkers for the "mail-in voting") is misleading or not



There is NO WAY (ZERO!) that Mail-In Ballots will be anything less than substantially fraudulent. Mail boxes will be robbed, ballots will be forged & even illegally printed out & fraudulently signed. The Governor of California is sending Ballots to millions of people, anyone.....

[Traduci post](#)

2:17 PM · 26 mag 2020

20.055 Repost 10.271 citazioni 82.831 Mi piace 1.420 segnalibri



1.420



ChatGPT

The text you provided appears to be a quote from former U.S. President Donald Trump. He made such statements on Twitter and in public remarks during the 2020 U.S. presidential election. It's important to note that the claims about widespread fraud in mail-in ballots were widely debunked and discredited by election officials, fact-checkers, and courts.

Numerous studies and audits found no evidence of widespread voter fraud in the 2020 election. The U.S. Department of Justice also found no evidence that would have affected the outcome of the election. The 2020 election was described by the U.S. Cybersecurity and Infrastructure Security Agency (CISA) as "the most secure in American history."

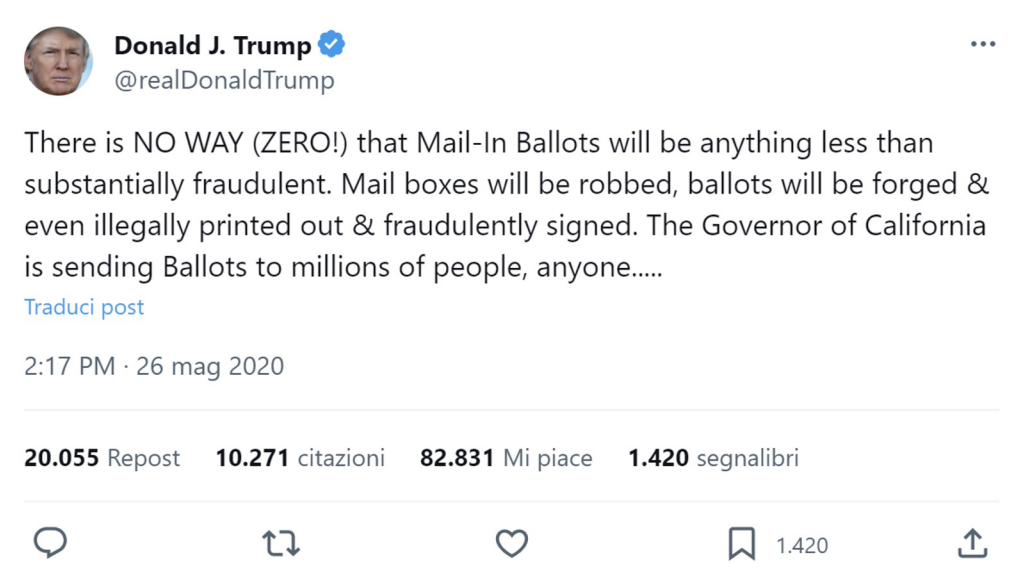
Therefore, the claims made in the statement you provided are not supported by credible evidence, and the assertion that mail-in ballots are substantially fraudulent is false. It's essential to rely on accurate information and verified sources when evaluating the legitimacy of statements, especially those related to elections and democracy.



For more: <https://nbcnews.com/politics/donald-trump/twitter-fact-checks-trump-s-misleading-tweet-mail-voting-n1215151>

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

Asking to ChatGPT whether a Trump's tweet (classified as misleading by fact-checkers for the "mail-in voting") is misleading or not



ChatGPT

The text you provided appears to be a quote from former U.S. President Donald Trump. He made such statements on Twitter and in public remarks during the 2020 U.S. presidential election. It's important to note that the claims about widespread fraud in mail-in ballots were widely debunked and discredited by election officials, fact-checkers, and courts.

Numerous studies and audits found no evidence of widespread voter fraud in the 2020 election. The U.S. Department of Justice also found no evidence that would have affected the outcome of the election. The 2020 election was described by the U.S. Cybersecurity and Infrastructure Security Agency (CISA) as "the most secure in American history."

Therefore, the claims made in the statement you provided are not supported by credible evidence, and the assertion that mail-in ballots are substantially fraudulent is false. It's essential to rely on accurate information and verified sources when evaluating the legitimacy of statements, especially those related to elections and democracy.



AI based Fact
Checking

For more: <https://nbcnews.com/politics/donald-trump/twitter-fact-checks-trump-s-misleading-tweet-mail-voting-n1215151>

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

Meta announced a large language model called LLaMA 2

A collection of pre-trained and fine-tuned large language model ranging in scale from 7 billion to 70 billion parameters

It did not release the model publicly but allowed researchers *to request access through a form*

Since it uses the PEFT/LoRA approach for fine-tuning, it decreases the computational and storage costs

Introducing
Llama 2



Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

Bohdan M. Pavlyshenko, in his work, basing on the open-source customized fine-tuning phase in order to enhance the detection of manipulated contents in the context of Information Disorder

- **PROMPT** used to give instructions and text to the model, make questions, etc.

```
<s>[INST] <<SYS>>
```

```
You are an expert in fake news and disinformation revealing  
and analytics.
```

```
<</SYS>>
```

```
{ prompt } [/INST]
```

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

PROMPT

```
<s>[INST] <<SYS>>
```

```
You are an expert in fake news and disinformation revealing and analytics.
```

```
<</SYS>>
```

```
Please analyse input text from perspective of revealing disinformation narratives.
```

```
Input text:
```

```
Hoping Ukraine will beat Russia "is not just a misunderstanding. It is a lie. It's impossible. Ukraine has no chance of winning the war against Russia. We missed the historic opportunity" to admit Ukraine to NATO, Russia is getting "stronger and stronger." Ukraine's admission to NATO "is not a realistic proposal at this moment, so forget about it".
```

```
"Call back Trump! Because you know, you can criticize him for many reasons ... but ... the best foreign policy of the recent several decades belongs to him. He did not initiate any new war, he treated nicely the North Koreans, and Russia and even the Chinese ... and if he would have been the president at the moment of the Russian invasion [of Ukraine], it would be not possible to do that by the Russians."
```

```
"Trump is the man who can save the Western world" and all of humanity. [/INST]
```

QUESTION

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

Given a question (the input)

The first narrative in the input text is that Ukraine cannot win the war against Russia. This narrative is manipulative as it suggests that Ukraine is incapable of success, which is a subjective claim not based on factual evidence.

The second narrative suggests that Russia missed a historic opportunity to admit Ukraine to NATO. This narrative is misleading as it oversimplifies the complex geopolitical situation and ignores the fact that Ukraine's admission to NATO requires a consensus from all member states, which is difficult to achieve due to various reasons.

The third narrative suggests that the US President, Donald Trump, is the only savior who can save the Western world and humanity from Russia. This is manipulative as it uses an emotional appeal to sway readers' opinion, ignoring the fact that Trump's policies are controversial and have been criticized by many.

LLama 2 detection provides the following results, first and third narratives are **manipulative** while the second narrative is **misleading**

Role of AI in detecting Computational Propaganda and Cognitive Warfare weapons

The efficiency of LLaMA 2 also helps in identifying

- the offensive lexicon in a text,
- what are the words and the named entities used to create a negative, neutral or positive sentiment,
- the tone, and the way in which a person is manipulated by a content

The offensive lexicon used in the text includes terms like 'lie', 'missed', 'stronger', 'foreign policy', 'initiate', 'treated', 'man who can save the Western world', 'humanity'. These words are used to create a negative sentiment towards Ukraine, NATO and the US, while glorifying Russia.

The named entities associated with negative sentiment in the text include Ukraine, NATO, and Trump, while Russia is associated with a positive sentiment.

The overall tone of the text is manipulative, using emotional language and subjective claims to shape readers' opinion. It manipulates the readers' perceptions by presenting a one-sided view of the situation, which is not based on factual evidence.



International & European
Criminal Law Observatory



UNIVERSITÀ DEGLI STUDI
DI SALERNO



Co-funded by
the European Union

Aspects and challenges of AI in Information Disorder detection

AI relevant aspects and challenges in ID detection

Using Artificial Intelligence in the detection of Information Disorder implies a large amount of advantages such as

- Improved speed, scalability and efficiency, by analyzing larger amounts of data (from different sources), detection and timeliness in response are better
- Machine Learning-based systems and Natural Language Processing (NLP) algorithms development that enable identification of patterns and trends in dissemination
- Multichanneling, by monitoring a wide range of channels (social media, websites, forums, etc.), we seek to understand the path of diffusion
- etc.

AI relevant aspects and challenges in ID detection

However, it also implies challenges such as

- **Presence of bias in data training**, machine learning algorithms are only as good as the data they are trained on. If the data are biased or incomplete, the algorithm may not be accurate in detecting disorder
- **Complexity** of natural language
- **Evolution of information manipulation tactics**
- **Ethical concerns**, since algorithms require large amounts of data collected from social network platforms, they include personal information about individuals. So, the efforts to identify and counter information disorder must be balanced with the protection of user privacy

AI Safety Summit

"As artificial intelligence rapidly advances, so do the opportunities and the risks"

United Kingdom hosted the first global AI Safety Summit in **1-2 November 2023**, bringing together leading AI nations, technology companies, researchers, and civil society groups to turbocharge action on the safe and responsible development of frontier AI around the world



AI Safety Summit

Considering

- the *necessity* of an immediate action to build a shared understanding of frontier AI,
- the *need* for an inclusive approach to address frontier AI and other risks,
- the *importance* of addressing current AI risks alongside those at the frontier,
- the *value* of appropriate standardisation and interoperability in AI and
- the *need* to develop the broader AI ecosystem including skills and talent

there is the need to have

- 1) a shared understanding of the risks posed by frontier AI,
- 2) a forward process for international collaboration on frontier AI safety, including how best to support national and international frameworks
- 3) appropriate measures which individual and organisations should take to increase frontier AI safety