

# Visual Generative AI in Warfare and Terrorism

## Risk Mitigation through Technical Requirements and Regulatory Insights

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieurin**

im Rahmen des Studiums

**Software Engineering & Internet Computing**

eingereicht von

**Tahel Singer, BSc.**  
Matrikelnummer 11740964

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Associate Prof. Dipl.-Ing.in Dr.in techn. Hilda Telloğlu

Mitwirkung: Dr. Kevin Marc Blasiak

Dr. Myriam Dunn Cavelty (extrenal supervision, ETH Zürich)

Wien, 21. Oktober 2024

---

Tahel Singer

---

Hilda Telloğlu



# Visual Generative AI in Warfare and Terrorism

## Risk Mitigation through Technical Requirements and Regulatory Insights

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieurin**

in

**Software Engineering & Internet Computing**

by

**Tahel Singer, BSc.**

Registration Number 11740964

to the Faculty of Informatics

at the TU Wien

Advisor: Associate Prof. Dipl.-Ing.in Dr.in techn. Hilda Telliöğlu

Assistance: Dr. Kevin Marc Blasiak

Dr. Myriam Dunn Cavelty (extrenal supervision, ETH Zürich)

Vienna, October 21, 2024

---

Tahel Singer

---

Hilda Telliöğlu



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Erklärung zur Verfassung der Arbeit

Tahel Singer, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 21. Oktober 2024

---

Tahel Singer



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Acknowledgements

First and foremost, I would like to warmly thank Prof. Hilda Tellioglu for providing support throughout both my Bachelor's and Master's studies, and especially the writing of this thesis. From my first day at university, she has been a role model for me, with her inspiring professionalism and personal dedication, always lending an open ear to her students. Thank you for your encouragement and for letting me write this thesis through such an exceptional collaboration.

I would like to sincerely thank Dr. Kevin Blasiak for guiding and supporting me throughout the entire process of writing, leading to new ideas, offering valuable insights, and assisting me in every matter. Your guidance gave me invaluable help and played a huge part in this work. Your feedback pushed me constantly toward excellence which I greatly appreciate and made this journey both fulfilling and rewarding.

Moreover, my deepest gratitude goes to Dr. Myriam Dunn Cavelty, without whom this work would not have been possible. She played an integral role in shaping the initial ideas and steps, establishing a clear research direction, and continuously offering her helpful feedback. I am truly grateful for the opportunity to engage in high-quality research under her extraordinary guidance, and for her patience while I navigated such a complex topic. Her office was always open to me, providing a space for hours of enriching discussions that not only deepened my understanding but also inspired my intellectual growth.

Furthermore, I would like to thank all of my closest and dearest friends whom I met throughout my academic path, in Vienna, Lausanne, and Zurich. Thank you for the support throughout the years, for growing together, and for teaching me that a physical distance has no meaning when the hearts are so closely connected.

Last but not least, I want to thank my family for their continuous belief in me and their endless love and support. Thanks to you I have been able to pursue my own path in every area I desired, gain invaluable life skills, and cultivate critical thinking from a young age. I could not have asked for stronger role models than you.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Abstract

The nature of modern terrorism and warfare has evolved significantly, with technological advancements enabling the capture and dissemination of uncensored graphic propaganda across social media platforms. These visuals, often in HD quality, can be used to train generative AI models, raising concerns about the misuse of such technologies to fuel violence, radicalization, and polarization, with profound psychological consequences on both micro- and macro-levels. This thesis examines whether current AI regulations, particularly the EU AI Act, adequately address these risks, and seeks for relevant technical solutions to mitigate them. We built a working corpus using the PRISMA framework, drawing on research addressing AI-powered radicalization and online terrorist activities. Through a socio-technical lens, we explored how exposure to violent content triggers radicalization pathways, studying radicalization models and the interplay between structured online and offline terrorist activities. We also explored the role of internet infrastructure and core algorithms in facilitating radicalization and how extremist groups exploit these social and technical components to achieve their goals, leading to a broad scope of direct and indirect consequences. Our analysis of the risk landscape, based on a risk-based approach, identified multiple risks, including propaganda-driven dehumanization, the enhancement of the “othering” phenomenon, the normalization of violence, and widespread psychological harm. We conducted a gap assessment of the EU AI Act, finding that while the act broadly covers these risks, it addresses key challenges like bias, privacy, transparency, and explainability only in abstract terms, without explicit technology-focused requirements. Additionally, there is insufficient focus on extremist groups and terror organizations as malicious actors, limited technological standardization, and no national education programs to build resilience against the misuse of generative AI. We recommend incorporating systematic human moderation, advanced machine learning algorithms to detect extremist inputs and violent outputs, and anonymization of individual visual attributes using generative adversarial networks (GANs). Furthermore, we propose a set of standards for watermarking techniques to support global regulatory efforts and research. These gaps highlight the need for active collaboration among regulators and other stakeholders to ensure the responsible development and deployment of AI technologies that mitigate the risks identified in this work.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Contents

<b>Abstract</b>	<b>ix</b>
<b>Contents</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Aim of the Work . . . . .	2
1.3 Structure of the Work . . . . .	4
<b>2 Background</b>	<b>5</b>
2.1 Generative AI . . . . .	6
2.2 AI Regulation . . . . .	14
2.3 Risks of Generative AI . . . . .	21
<b>3 Methodology</b>	<b>27</b>
<b>4 Analysis</b>	<b>31</b>
4.1 Socio-Technical Systems . . . . .	32
4.2 PEOPLE: Individuals and Radicalization Processes . . . . .	34
4.3 TECHNOLOGY: Infrastructure and Radicalization . . . . .	38
4.4 STRUCTURE: Social Media and Terrorism . . . . .	49
4.5 TASKS: Propaganda, Polarization, and Psychological Harm . . . . .	53
<b>5 Results</b>	<b>67</b>
5.1 Primary Risks of Generative AI . . . . .	68
5.2 EU AI Act Gap Assessment . . . . .	85
5.3 Technical Requirements for Risk Mitigation . . . . .	94
<b>6 Conclusion</b>	<b>105</b>
<b>List of Figures</b>	<b>109</b>
<b>List of Tables</b>	<b>109</b>
<b>Bibliography</b>	<b>111</b>
	xi



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Introduction

## 1.1 Motivation

Graphic violent content has reached a disturbing new level, broadly present in recent warfare, terrorism, and the outputs of generative AI models. These significant changes appear to co-evolve and bidirectionally influence each other in a concerning direction that escalates the cycle of increasing violence. As AI technologies make huge steps by leaps and bounds toward generating hyperrealistic visual content, we start acknowledging the full scope of the potential risks encompassed in the combination of available real-life extreme visual violence and Generative AI models. The urgent need to contain them is hence also a rising concern. The technological capability of AI models to generate such high-quality visual content significantly exacerbates the challenges that the rising post-truth era poses, coupled with radicalization tendencies and societal polarization. In this work, we intend to trace and modulate the pathway that exposure to graphic violence triggers, in particular in the context of war and terrorism, and extract the concerning risks along the way.

Terrorism's intrinsic goal is to instill fear and sow psychological distress among large populations in the name of an ideology, beyond the immediate physical harm. For this cause, AI is a great technological tool that terror organizations can easily misuse to harm individuals on a large scale. Despite the factual statistics that terrorist attacks are often less deadly than other tragedies and traumatic occurrences, e.g. car accidents, extensive media coverage can significantly amplify their psychological impact and escalate fears globally, Kuey [96] thoroughly discusses. Due to the extensive media coverage, both through traditional journalism and the rise of social media, acts of terrorism and the manner in which warfare takes place have long left the sheer physical dimension.

The appearance and virality of human-made horrors have undergone a significant transformation. The horrible beheading videos of ISIS starting in the year 2014 marked the

beginning of a whole new era of visibility and “aesthetics” of violence targeting the large global crowd, as Friis [59] addresses. Further horrific cases occurred over the years, including the first fully live-streamed massacre in Christchurch, New Zealand, in 2019; the terror attack in Vienna in November 2020 that was caught by cameras of surrounding civilians; the Ukraine-Russia war; and the ongoing conflict zone in Israel-Gaza, particularly since the October 7th attack.

The public and the government’s interest in disseminating graphic content differentiates drastically from one culture to another and governments have followed different censorship policies. These significant differences demonstrate as well the difficulties in collectively addressing the rising global problem of dealing with viral extreme graphic violence. Regardless of the approach taken, broad exposure to visual violence raised major concerns regarding national mental health among specialists. For instance, the Ministry of National Security for Child Protection in the cyberspace of Israel published a national warning statement in Hebrew<sup>1</sup> and Arabic<sup>2</sup> following the viral graphic content of abusing people and corps that spread from the October 7th attack [74].

The whole new pool of high-quality, graphic content from the above-mentioned real-life atrocities can serve as a training dataset for generative AI models, facilitating the creation of increasingly realistic and disturbing fictional violence and brutalities. Unlike horror scenes from the movie industry, filmed terror attacks capture genuine moments of human suffering while promoting an extremist agenda for propaganda purposes. The authenticity of real events can result in well-nuanced outputs by AI systems. There is no guarantee that the outputs from those machines would not be disastrous, particularly due to the lack of strict and unified restrictions about admissible content. As AI systems evolve and their capabilities become apparent along with their running time, we learn about the various actors and their interests in perpetuating and amplifying violent ideologies through violent content. The phenomena of radicalization and polarization have dominated the latest conflicts and terrorist activities. The immersive nature of high-quality visual content has fueled the emotions of individuals, either engaging them deeply in occurrences or causing mental health harm.

### 1.2 Aim of the Work

This research aims to explore and name the whole new unknown series of risks that the convergence of recent historical-political events and technological advancements poses in this evolving context. Subsequently, based on the discovered risks, we also aim to emphasize the increasing need to adopt new regulatory approaches and suggest explicit technical requirements for the existing legal framework. Therefore, our research question is a compound of three parts:

---

<sup>1</sup>[https://www.gov.il/he/departments/news/warning\\_105](https://www.gov.il/he/departments/news/warning_105), Accessed on: 17.03.2024

<sup>2</sup>[https://www.gov.il/ar/departments/news/warning\\_105](https://www.gov.il/ar/departments/news/warning_105), Accessed on: 17.03.2024

- RQ1:** What are the primary risks posed by visual generative AI in exacerbating radicalization, polarization, and psychological harm during times of war and extreme violence?
- RQ2:** What regulatory and technology-specific gaps exist in the current framework concerning the identified risks?
- RQ3:** What technological solutions should be incorporated into future regulations, and how do they mitigate the identified risks?

We seek to clarify how malicious actors, with a focus on terrorism and extremism, can exploit AI technologies to target prone recruits and propagate their violent ideologies. The regulative gaps aim to highlight both policy aspects and necessary technological requirements. Technological requirements offer solutions to address the identified risks and fill the technology-specific gaps in the existing legal framework. Our findings regarding technological requirements aggregate knowledge acquired from the literature review in the background and our working PRISMA corpus, considering the context of the raised risks. Explicit technological requirements in regulation is crucial to enhance the resilience of generative AI models against misuse by extremist actors, minimizing their risk of being weaponized for harmful purposes, and ensuring a secure digital environment.

We use a method mix to comprehensively answer the two parts of our research questions, which include both quantitative and qualitative methods. We first build a corpus for the literature review following the PRISMA framework<sup>3</sup> to obtain a cross-modality overview with which we can address the referred main risks in the first part of our research question (RQ1). We then conduct a document analysis of the EU AI Act to answer our second part of the research question (RQ2) with which we create a gap assessment to identify the missing technical requirements. Finally, we address the missing technological requirements by discussing adequate technological solutions that can mitigate the identified risks (RQ3).

Throughout our research, some leading aspects guide us. These include the psychological impact of exposure to violence, its potential to normalize harmful behaviors, and the risk of radicalization through accessible and generated violent content. We acknowledge that addressing this global challenge requires a coordinated response from governments, international organizations, and technology companies following democratic values as the leading tone. Therefore, this work targets these stakeholders and wishes to facilitate a common knowledge base for varied domains.

---

<sup>3</sup>PRISMA stands for *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*. This structured approach consists of a set of guidelines designed for conducting systematic reviews and meta-analyses. This framework is mostly present in health research but can also be applied across various disciplines [104].

### 1.3 Structure of the Work

The work encompasses a theoretical background (chapter 2), the methodology (chapter 3), an analysis of the constructed corpus (chapter 4), the results (chapter 5), and the conclusion including future work (chapter 6).

The background in chapter 2 would guide us through the technological basics of generative AI, the historical course of AI regulation, the current legal state-of-the-art focusing on the European sector, and the already known risks of generative AI that are relevant to our context and scope.

We construct our working corpus and define the key categories of our legal document analysis in chapter 3. The systematic method of PRISMA allows us to create a scientific basis for our analysis which is done in the following chapter. As part of our analysis in chapter 4, we apply a socio-technical lens to understand better the interconnections between the technological and social components.

The results in chapter 5 portray the risk landscape of our researched context. The first part of our results draws a clear line between the known risks presented in the background and the findings of our analysis along with their nuances. The second part systematically investigates the chosen categories and identifies the gaps in addressing them in the EU AI Act. The third part synthesizes the acquired technical knowledge throughout the thesis and provides technical solutions to mitigate the identified risks.

Finally, we conclude our findings in chapter 6 and refer to the limitations of the work and open points for future research.

# CHAPTER 2

## Background

The application of Generative AI in the context of atrocities, including acts of terrorism, armed conflicts, and wars, can lead to hyperrealistic visual representations that used to be far from imagination. This technological jump in the current state-of-the-art has become feasible thanks to many developments in the last couple of decades, with a marked acceleration in recent years. These include big data, cloud computing, powerful hardware, and new methods to develop and train AI models. The large pool of visuals from horrors occurring throughout human history has evolved majorly; due to the technological infrastructure that captures those moments in HD quality and stores them in large-scale databases.

Within our definition of visual violence falls the term “media violence”, which Siddika [168] defines as “any visual portrayal of violent physical activity or violent thought by one person or character against another”. Huesmann et al. [78] add and specify that media violence is merely exercised in a non-physical dimension and strictly stays within the screen. Along with the evolution of offline an online violence, which has become more visual, accessible, and viral, we gradually discover further related aspects.

Of particular concern are the ethical questions surrounding the risks that this new form of violence and the following shifting norms and values of our globalized society pose. The next stage in discussions about ethics and norms often includes the concerning local and global regulative state and how up-to-date they remain, especially in a constantly evolving technological landscape. The potential for AI-generated visual violence is entangled within the chain of technological systems, the associated risks they present, and the evolving regulation. The dynamic interplay between the technological component, namely, the generative AI models, the regulative component, including the ethical norms and the regulation, and emerging risks is constantly evolving. A comprehensive understanding of these components is thus essential for research into potential radicalization powered by AI.

Throughout the background chapter, we learn how the latest generative AI techniques efficiently produce diverse outputs that are highly similar to their training data which plays a significant role in reproducing violent visuals. We also discuss the most relevant technical challenges of generative AI systems and develop a solid understanding of watermarking techniques, which have become a regulatory standard concerning visual generative AI systems. We then see how the established common ethical AI values are embedded into the comprehensive AI regulation framework of the EU. As the most recent legal state-of-the-art, the EU AI Act aligns with current best practices and has the potential to ensure global cohesion for future legislation. Its proposed risk-based approach will accompany this work and be the reference point for improvement suggestions. Our concluding literature review about the main researched risks of generative AI emphasizes the manipulation of information as a primary risk, especially linked to radicalization and polarization purposes. These found risks serve as the core knowledge and model of reference to our work. We explore the risks concerning visual AI in the context of extreme violence, as seen in war and terrorism. This focus is currently of special importance, as the focus of research has shifted toward understanding the rising tendency of perpetuation of graphic violence in generative AI systems only recently. With the help of the most relevant technical, regulative, and societal aspects brought in the background, we can proceed with studying the under-explored risks of our domain profoundly and find bridges to their potential mitigation in the chosen EU AI act and technical solutions.

### 2.1 Generative AI

Generative AI stands for artificial intelligence systems that automatically generate new textual, visual, or audio content based on provided training data and multi-layered algorithms. Cao et al. [30] thoroughly examine the evolution of generative AI technologies. They explain that Generative AI accelerates the process of creating content, which has become more efficient, quick, and easily accessible resulting in high-quality content. Generative AI models offer outputs for different tasks, including text-to-text, text-to-image, text-to-video, and text-to-audio, operating in both unimodal and multimodal forms. The latter integrates different types of tasks into one model and generates different types of outputs. Unlike the older models, which were segmented by tasks, this allows for more general use with a cross-modality characteristic. Generative AI models and techniques bring the technological capabilities of traditional AI to the next level through their increased size of datasets and computational power. Beyond the technological developments, generative AI systems suffer from multiple inherited problems, including bias, factual inaccuracies, data privacy, transparency, and explainability. There are ongoing efforts concerning detection techniques of generated content, with the leading watermarking mechanism that has become a legal standard.

### 2.1.1 The Evolution and State-of-the-Art of Generative AI

Traditional AI mostly differs from Generative AI in its requirement for less training data, algorithmic approach, and generally lower quality of the delivered output. The first AI systems were rule-based with pre-defined rules and logic, being able to cover only narrow domains with generalization abilities. Later AI systems often incorporated solely statistical machine learning techniques and one-layered neural networks with a limited capability to process information and generate higher complexity outputs. Later, the deep learning approach took over, including one or more layers of neural networks that aimed at imitating the brain structure. These AI algorithms are highly task-dependent and their models are accordingly adapted. Traditional AI at its early beginning was transparent and explainable and along its development, the systems have more components and layers and the outputs become less explainable and traceable, becoming “black box” systems [30].

A further step toward AI that generates content came with the transformer architecture, first introduced in 2017 by Vaswani et al. [180]. A new feature of this approach is its multimodality, namely, one encoder-decoder transformer integrates different domains allowing one domain to benefit from the knowledge of the other, as the training data includes different types of data. An additional recent approach named Generative Adversarial Networks (GANs) was introduced in 2020 by Goodfellow et al. [66]. In this approach, two neural networks, the generator and the discriminator, are trained simultaneously such that the generator repeatedly creates synthetic data samples, and the discriminator evaluates their authenticity against non-synthetic data. With each iteration, the model reinforces the acquired knowledge which significantly improves the quality of the generated output with the repeating feedback which is called the adversarial process. Bengesi et al. [18] presented the architecture of GANs in their work 2.1.

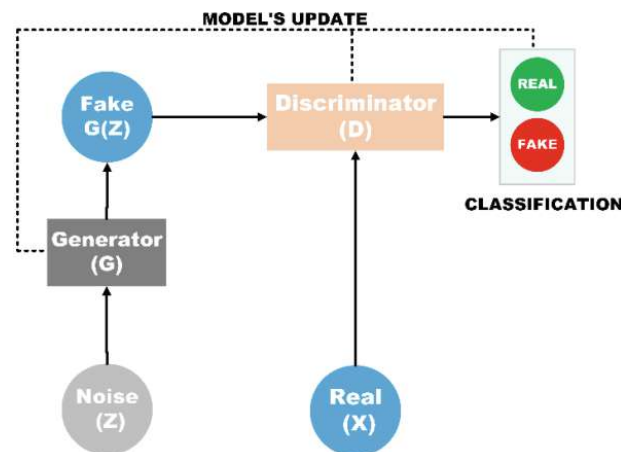


Figure 2.1: GAN architecture [18]

The generative AI tasks of text-to-image and text-to-video generation involve mul-

multiple stages, including the creation of new visuals or the augmentation of existing ones. These processes begin with the handling of textual or audio input with an optional visual input and proceed to deliver a relevant output based on the extracted information. The methods used can vary across models, incorporating both supervised and unsupervised machine learning techniques [30]. Models with supervised learning are trained on labeled input-output datasets that human annotators curate. This human oversight helps ensure a reliable, accurate, and fair ground truth, on which the model bases its future predictions. In contrast, unsupervised learning works with unlabeled data and without the explicit intervention or guidance of humans in the process. This gives a free hand to the model to independently identify patterns in the dataset without any enforced ethical guidelines. However, also in unsupervised learning, it is possible to keep some human oversight with post-processing through manual interpretation of the results, their validation, and refinements in the learning process, such as adjusting the relevant parameters, or the selected features. Human intervention is also referred to as “humans in the middle” or “humans-in-the-loop” and is essential for the mitigation and integration of agreed human ethical guidelines in AI systems. It aims to mitigate potential bias and increase the quality and reliability of the outputs [90].

Textual input is firstly handled mostly by tokenization, in which sentences are divided into smaller units, the “tokens”, that are the most optimal atomic units for language processing tasks. The tokens are then converted into numerical embeddings using different techniques, e.g. Word2Vec (converting words into high-dimensional vector representations). This conversion into embeddings enables the efficient capture of the semantics of the units within the context of the given input and helps understand linguistic nuances.

Audio input processing falls under the task of automatic speech recognition (ASR) and involves signal processing techniques that map raw audio waveforms into the correct string of words in two steps; first, transforming the input waveform into a sequence of acoustic feature vectors (the digital representation) and second, extracting meaningful information for each small time window of the signal<sup>1</sup> that finally results in the textual representation of the given audio with which it can be proceeded further similarly to the textual input of the generative AI pipeline.

Processing of visual input is often done with the deep learning architecture of convolutional neural networks (CNNs). They extract hierarchical features from the visuals, including edges, textures, and patterns, which enables the extraction of features and the recognition of structures and objects within images for image classification and object detection [90].

The technologies relevant to our research involve the latest releases of generative AI text-to-image and text-to-video models that create diverse and high-resolution outputs of visual harmony that often closely resemble the training data provided. These include

---

<sup>1</sup>The window size is defined by the length of the segment of the signal in milliseconds.

the main models for image generation by OpenAI's DALL-E<sup>2</sup>, Midjourney<sup>3</sup>, and Stable Diffusion<sup>4</sup>. The core architecture at the base of these models are either diffusion models or transformers.

DALL-E's architecture is transformer-based. Transformers are part of the deep neural network approach. Transformers do not process data sequentially like earlier methods, e.g. recurrent neural networks (RNNs) and convolutional neural networks (CNNs), and offer a decent solution for problems concerning large inputs and outputs (gradient explosion, long computation of local and global minima, etc.). They integrate a self-attention mechanism with which the model evaluates the importance of each of the parts of the input sequence within a context. In addition, they employ multi-head attention to compute parallel, and a word embedding layer<sup>5</sup>. Thanks to these concepts, transformers can capture contextual relationships together with the long-range dependencies in an effective way and compute parallel [18].

Midjourney and Stable Diffusion use diffusion models. These were introduced after the transformer and include three main formulations; denoising diffusion probabilistic models, stochastic differential equations, and score-based generative models. They consist of two-step forward and reverse diffusion processes. The first step employs (Gaussian) noise to the training data which alters it entirely. The second step, the denoising, iteratively reverses the added noise and creates eventually new data that is highly similar to the original data. This way, the model learns how to generate diverse and high-quality outputs that are suitable for a large scope of tasks [18].

Another newest development that is also a big player in the field is OpenAI's video model, Sora<sup>6</sup>, which produces up to one-minute-long video scenes from text instructions. This model was first announced on February 15, 2024, and as of now (last stand August 2024) has not been released to the public yet. This has the trivial implication that the limitations and edge cases of this model are only now to be discovered. For instance, in an article on "der Standard" from March 2024 [2], the concern regarding the production of videos involving nudes was raised. Moreover, the current context of the ongoing wars, both between Russia and Ukraine and in the conflict zone of Israel and Gaza, is yet to be fully studied. Sora leaves behind other more limited AI video generation models, namely, Synthesia<sup>7</sup> which generates AI videos with avatar figures and Meta's Make-A-Video that Singer et al. [169] introduced. OpenAI's newest model manages to keep a visual consistency through the whole frames with complex and detailed characters.

<sup>2</sup>Currently, in its third version <https://openai.com/index/dall-e-3/>, accessed on 23.08.2024

<sup>3</sup><https://www.midjourney.com/home>, accessed on: 23.08.2024

<sup>4</sup><https://stablediffusionweb.com/>, accessed on: 23.08.2024

<sup>5</sup>Word embedding is a technique in language processing in which words or tokens are represented as continuous vectors in a high-dimensional space. This way, they capture semantic relationships within a corpus and take advantage of similar meanings.

<sup>6</sup>[openai.com/sora](https://openai.com/sora), accessed on: 24.03.2024

<sup>7</sup>[synthesia.io](https://synthesia.io), accessed on: 24.03.2024

On a technical note, Sora is inherently based on a diffusion transformer. The initial step is compressing videos into a lower-dimensional latent space, a technique that is called spacetime latent patches<sup>8</sup> and uses the patches as the building blocks of the compressed video. With a similar concept to tokenization in natural language processing, each of the patches provides the main model with visual phrases that are then used in the process of constructing the output videos. The initial frame is filled with visual noise, according to the diffusion transformer’s pipeline, and then step-wise the model denoises the image until the generated video emerges, as presented in the technical report of Sora [133] in 2.2.



Figure 2.2: Sora’s stepwise denoising process on the sequence of the patches [133]

However, the currently available technical report does not disclose specific information regarding the training methodology of the video captioner that produces the video descriptions [105].

### 2.1.2 Technological Challenges of Generative AI

Nevertheless, generative AI systems amplify already known, inherited problems that AI systems have posed, which Goktas [63] addresses in his latest comprehensive bibliometric work. The primary technical challenges consist of inherent bias, the generation of incorrect information, and privacy breaches, which technically necessitate increased transparency and explainability of these systems. There is an increasing interest among researchers to understand the core morality of AI systems and identify “artificial evil” tendencies, as Perov and Perova [138] noted. We address later in this chapter the ethical considerations and risks that generative AI technologies pose.

Bias in technology manifests as systematic decisions and outputs that arise from prejudices present in the training data or algorithms, potentially resulting in errors or unfair outputs. Skewed or unrepresentative training data is then reflected during the model’s decisions and final outputs, leading to a model with a narrow set of known patterns and structures. A simple example is the classification task, for which a model fails

<sup>8</sup>The learned representations within a neural network capture the spatial features of the image or frame and the changed dynamics over time of the video. The patches help models process and understand visual information in motion by identifying crucial patterns across space and time.

to identify an object that is not present in the training data. Algorithmic bias can still persist even with diverse training datasets, due to embedded pre-selected features and attributes, potentially enforcing unfairness or injustice. A trivial example is an algorithm for university admissions in which only boys are selected, excluding women from consideration [63].

The technological term for the tendency of incorrect information is hallucinations, which are significant deviations from the facts, factual inaccuracies, or common knowledge. Hallucinations can be textual but also visual. Textual deviations can include textual content that entirely contradicts well-established facts, for instance, naming wrong information regarding the term of political figures or inventing stories that do not appear in the bible. Visual deviations are for example clearly unrealistic or distorted body parts of people or animals. These cases happen due to insufficient diversity and pre-existing bias in the data training or by wrong assumptions that the model makes. This phenomenon is especially worrying because the outputs give the impression to be highly credible and realistic and hence the precision of the outputs is vital for the security aspects of this technology. Images and videos affected by hallucinations can appear as extremely trustworthy, and disseminate scenarios of completely fake events and false narratives [138].

Data privacy and potential breaches arise from the incorporation of sensitive data into training datasets, leading to outputs that replicate private belonging to natural people and organizations. Sensitive data includes personally identifiable information (PII) that has not been anonymized and is present as unmasked plain information in datasets. Such outputs can harm privacy, resulting in revealing confidential information and enabling unauthorized disclosures. The issue of consent is crucial, especially given that much of the data is harvested from social media platforms, where users frequently and extensively share personal details. This practice could yield outputs that closely resemble real individuals in fake scenarios, mimicking their way of writing styles or physical attributes [63].

Transparency regarding the operation of generative AI models, particularly concerning the training datasets and decision-making processes, remains often opaque. Generative AI systems involve complex architectures, including deep learning and neural networks, and incorporate cutting-edge techniques and advancements that are undisclosed due to the competitive market. Furthermore, The vast amounts of training data can be too extensive for clear oversight. The transparency of AI systems is key to establishing trust among users and stakeholders. It can contribute to the identification of biases or errors and enhance the accountability of the systems, which is also essential for their assessment [63]. Increased transparency could, however, also lead to increased risks, such as cloning models and manipulation by malicious parties [138].

Explainability, which is a closely related term to transparency, is the technical ability to track back the steps leading to a particular output, contributing to the robustness of AI systems, and the identification of biases or other flaws in them. The chain of

decision-making remains often obscure in generative AI, due to their complex nature, multiple steps, and large training datasets. The large-scale unannotated datasets contribute to the explainability challenge, especially the inability to identify the source of bias or inaccuracies of the output [63].

The five primary technological challenges in generative AI systems - bias, misinformation and disinformation, privacy, transparency, and explainability - will be central to our analysis. These five challenges will significantly influence our results, guiding both the evaluation of regulatory gaps and the discussion of effective technological solutions.

### 2.1.3 Digital Watermarking

To enhance the accountability of generative AI systems and the trust of digital content, while also addressing their misuse potential, technical counter-measurements are in the development focusing on detection. These primarily include advanced AI algorithms for the detection of extremist use with a focus on hate speech, which we will discuss later in 4.3.3, metadata analysis, and diverse detection algorithms to verify whether a sample of data is synthetic or authentic. These include approaches such as deep learning-based classifiers and anomaly detection techniques, as Lu and Ebrahimi [108] thoroughly review and evaluate, but also digital signatures, called “watermarking”, that are incorporated into the outputs by the generating model. The aim of these detection mechanisms is to create a safer digital environment and mitigate the risks associated with generative AI misuse [70].

Watermarking is becoming a technologically inseparable global standard of generative AI models, in particular in the EU, as we will see later in 2.2.2, and the common state-of-the-art to identify generated visual input. Watermarks are in essence a few inverted bits in the image that are machine-readable. They can be injected after the image generation, called the “post-processing watermark” or either throughout the image generation process, called the “in-processing watermark”. Post-processing watermarks are model-independent and typically result in poorer image quality than the in-processing ones [44].

Watermarks were originally thought for protecting intellectual property of creators and are largely used for industry-specific needs, as Takale et al. [175] explore and therefore have to be customizable, flexible, scalable, easily integrable, while also ensuring security (utilizing encryption for data privacy and protecting against unauthorized image manipulation), traceability of the content’s origin, and providing a good performance concerning the algorithms efficient and preserving the perceptual image quality. Nagai et al. [126] offered a framework for effective watermarking algorithms for the image domain. Their primary requirements include:

- **Fidelity:** the visual quality of the image should not be decreased with the injection of the watermark.

- **Robustness:** common signal processing operations should not be capable of corrupting the watermark.
- **Capacity:** the ability to process a large amount of information effectively.
- **Security:** the watermark should be kept as a secret and keep security principles concerning unauthorized access, read, and modification. The secrets include the location of the watermark and the regularized parameters that serve as a secret key.
- **Efficiency:** the ability to run efficiently the watermarking process.

Nagai et al. refer to the overwriting of watermarks as a severe attack and emphasize the importance of creating a resilient watermarking mechanism. To overwrite successfully the visual watermark, the attackers have to know the location of the injected watermark. Ding et al. [44] mention additional distortion attacks, namely, geometric distortion (rotation, resizing, and cropping), photometric distortion (changes in the coloring definition of the image), degradation distortions (blur, noise, and compression), and the combination of all of these attacks.

Ding et al. [44] also examined adversarial attacks that have been quite successful. The embedding attacks include techniques that disturb the detection process of the watermark from proceeding by manipulating the image aiming to minimize, alter, or even remove the injected watermark. Unlike the sheer overwriting of the watermark, this attack is especially successful, because it manages to avoid the degrading in quality of the original image by alternating the encoding of the watermark which are non-visible characteristics of the content. The second type of adversarial attacks is the surrogate detector. In this attack, a decoder model is trained on a subset of known watermarks to infer and detect watermarks and verify the decoded messages embedded into the watermarked images. This model inherently mimics the behavior of a legitimate decoder of a watermark detector. This way, it can erase watermarks from the marked image or transfer the content to a plain new image omitting the watermark.

Hwang and Oh [82] summarized the currently offered watermarks of the leading technology companies. DALLE-E 2's watermarks are injected in five primary colors in the lower-right corner of the outputs. Microsoft injects a small b (for "Binge") on the lower-left corner of its visual outputs. It seems that the current proposed watermarking techniques by the market referred to by Hwang and Oh [82] do not meet the requirements set by Nagai et al. [126] and are highly vulnerable to the above-mentioned attacks, particularly due to known patterns and location. Software to remove watermarks is also available<sup>9</sup>.

<sup>9</sup>There is an offered app on Chrome Web Store for removing the watermark from DALL-E 2: <https://chromewebstore.google.com/detail/remove-dalle-2-watermark/gbebakhjjocdddnkkoaglcldpklmcmchi>, accessed on: 16.10.2024

The search for sufficient and secure watermarking for the image domain is still ongoing. There are no global standards governing the admissible algorithms, nor for the authorized entities or organizations that have access to the detection tools that verify the authenticity of the image. The current solutions and implementations are vulnerable to a broad attacks vector. This will need to change soon to meet the legal criteria, that we will explore in 2.2.2.

## 2.2 AI Regulation

Different components regarding the nature of AI technologies result in new social and science-oriented norms. Some of them include the high correlation between training data and algorithms to the increasing quality of the outputs as well as what core values they promote and the fact that complicated tasks that usually require human intelligence can be automatically performed. Alongside these changes, the ethical line of thought rooted in these systems has to reflect and maintain human and democratic values, especially as they have the potential to become a substantial element in the ecosystem of technology and society. Many players at the macro and micro levels have occupied with this understanding, starting with companies, continuing with the national-governmental level, and reaching the highest international level of international organizations, like the European Union, the United Nations, or OECD. The development and employment of AI technologies are global, while the leading values and priorities of the different actors can greatly differ which trivially leads to an inherent conflict. Part of this conflict's core questions include: who sets the tone, whose interests should be put at the front, what are the hard limits when it comes to innovation vs. protection of basic human rights and security, and what principles those systems have to follow and fulfill [87].

### 2.2.1 AI Ethical Guidelines

Generally, ethical guidelines establish a set of standardized principles concluding a multifaceted moral code of conduct for a specific domain that underpin regulatory frameworks. Their proposed fundamental values should lay the groundwork for legal documents while strengthening. Transparency about the values standing behind regulation fosters public trust in democratic processes and institutions. Ethical guidelines serve as a valuable tool for policymakers, helping them identify and mitigate potential issues and risks. Various stakeholders, including experts from the industry and academics, representatives of government and regulatory officials, non-profit organizations that advocate for the societal perspective or organizations from the industry that speak for the involved companies, contribute to their writing process. The diverse input integrating wide range of perspectives ensures that ethical guidelines are comprehensive and sound [87].

In the context of AI technologies, ethical guidelines aim to promote responsible development, implementation, and use of AI systems. Goktas [63] emphasizes the importance of ethical oversight in the context of generative AI technologies, in accordance with the

risks they pose. The publication of AI guidelines has been on the rise since 2016, mainly released in the United States, the European Union, and the United Kingdom. They mostly stem from private companies and governmental authorities, as Jobin et al. [87] mapped in 2019 and later Corrêa et al. [37] reconfirmed in 2023. Establishing ethical guidelines seems to be a continuously developing issue that concerns multiple stakeholders. It is of high importance to have an international baseline for the ethics of AI. One main concern is that governments can not only benefit from this technology but even exploit it under some circumstances. Furthermore, it is also crucial to have a local implementation of the guidelines at the company level, as they are the primary responsible stakeholder for developing and providing the AI systems that eventually serve both the private and public sectors. An external obligating framework can hinder and restrain the power of governments and force them to apply local rules that endorse secure development and employment. The significant milestones in the development of international AI ethical guidelines include the following documents, some of which have been updated over the years:

- **IEEE Global Initiative, Ethically Aligned Design, December 2016 [84]:** a breaking-through document that brought together hundreds of representatives from various disciplines and different continents to identify and provide recommendations for human-centred design of autonomous and intelligent systems that would finally align with the global IEEE standards, hence having a huge global impact.
- **EU Guidelines on Ethics in Artificial Intelligence, April 2019 [51]:** this document was a result of the European Parliament’s call to adjust the existing legal framework to fit a “human-centric” approach. This European Commission’s document established a non-binding ethical direction for legislative efforts by the EU and it aims at highlighting the required ethical considerations during the design, development, and implementation of AI systems. The EU is a key player in the global race on AI and its regulation. Its strategy toward regulating AI puts focus on the protection of human values and the resilience against social risks. The EU strategy differentiates from the US and China, while the first puts focus on private efforts and self-regulation and the latter is government-led.
- **OECD AI Principles, May 2019 [57][107]:** the OECD’s AI principles were adopted by the member states in 2019 and were once again updated in May 2024. They focus on trustworthiness and reliability of AI systems while balancing between their benefits and potential risks with specific recommendations for policy makers and AI developers. The economic growth is a key component in this document as well as the potential innovation and data governance. OECD has a significant role in establishing globalized policies frameworks and setting globalized standards; namely, the countries are expected to report back to OECD, mostly including their data and statistics, their policy implementation, and be a subject to peer reviews from other member states. OECD’s annual public reports that construct a public monitoring mechanism aim to foster best practices and enhance different

processes through healthy competition among member countries<sup>10,11</sup>. Nevertheless, the OECD-countries include only 38 members, leaving out many of the African countries, the Asian countries (only Israel, Japan, and South Korea), and Central and South America. Notable large and impactful countries that are not part of the OECD countries include China, Russia, India, Turkey, and Brazil<sup>12</sup>.

- **G20 AI Principles for responsible stewardship of Trustworthy AI, June 2019 [60]**: drawing from the OECD AI Principles, the G20 AI principles provide a higher-level governance principles for establishing a legal framework with a broader focus on global cooperation and economic implications<sup>13</sup>. The G20 is an international yearly summit comprises 19 countries (among them some of the non-OECD-countries, e.g. China, Russia, India, Turkey, and Brazil), the European Union, and the African Union, representing two-thirds of the world's population<sup>14</sup>. G20 has no hard enforcement mechanisms and is therefore only a soft power that enables a dialogue among member states and provides consensus-based and non-binding policies.
- **UNESCO's Recommendation on the Ethics of Artificial Intelligence, November 2021 [179]**: all of its 193 member states adopted the UNESCO's document on the ethics of AI in 2021, leaving a remarkable consensus on the subject and defining a clear framework for further local development. This document serves as an important reference point for the design of policies and regulations and reaches all the UN member states (besides Liechtenstein, Israel, and between the years of 2018 - 2023 also the US [31]) as well as three further non-UN members<sup>15</sup>. Similarly, also this suggested framework is non-binding and rather serves as a soft-power. This document puts at the front a broad the humanistic approach of prioritizing human rights and the societal influences.

The wide intersections between these ethical frameworks have been also quantitatively observed. Jobin et al. [87] found some main values among the first ethical guidelines until the year 2019 which Corrêa et al. [37] largely confirmed in their work from 2023. These main ethical values are shared by all our reviewed guidelines 2.1:

---

<sup>10</sup><https://www.oecd.org/en/about/how-we-work.html#set-standards>, accessed on: 04.08.2024

<sup>11</sup><https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0438>, accessed on: 04.08.2024

<sup>12</sup>[https://www.international.gc.ca/world-monde/international\\_relations-relations\\_internationales/oecd-ocde/index.aspx?lang=eng](https://www.international.gc.ca/world-monde/international_relations-relations_internationales/oecd-ocde/index.aspx?lang=eng), accessed on: 04.08.2024

<sup>13</sup><https://www.caidp.org/resources/g20/>, accessed on: 04.08.2024

<sup>14</sup><https://www.g20.in/en/about-g20/about-g20.html>, accessed on: 04.08.2024

<sup>15</sup><https://www.unesco.org/en/countries>, accessed on: 04.08.2024

Value	Meaning
<b>Transparency / Explainability</b>	promotes the disclosure of information, primarily about the use of AI the data use, as well as outputs that can be backtracked
<b>Trustworthiness / Reliability / Safety / Security</b>	emphasizes the consumers' trust in AI, the development process, and its applications
<b>Justice / Fairness / Non-discrimination</b>	promotes discrimination or outputs with unwanted rooted bias and forwards diversity, inclusion, and pluralism
<b>Beneficence / Non-maleficence</b>	promotes safety, security, human well-being, peace, and harm-prevention excluding specific high-risk applications that violate privacy, discriminate, or physically cause harm
<b>Responsibility / Accountability / Liability</b>	promotes the integrity of the development and employment of AI applications as well as aims to dissolve the ambiguity regarding legal ability concerns
<b>Privacy</b>	serves as both a standard and a right that has to be protected
<b>Freedom / Autonomy / Democratic Values</b>	promotes the freedom of expression and the freedom to flourish and self-determination
<b>Dignity / Human Rights</b>	promotes the development of AI systems that avoid harm, automated classification, or unknown human-AI interaction with emphasize on human rights respected by developers and supported by legislation
<b>Sustainability</b>	minimizes environmental damage while promoting AI systems with increased energy efficiency

Table 2.1: The main intersecting AI ethical values [87]

These values demonstrate a global commitment to ensuring the responsible and thoughtful development and use of AI technologies. They align with the technical challenges of generative AI systems discussed in 2.1.2, namely bias, reliability, privacy concerns, transparency, and explainability, and take into account their societal impacts, contributing to a more comprehensive understanding of human-centric AI. They provide a foundation for framing and creating binding regulations based on these shared principles. They give legitimacy to regulations and serve as an important tool for citizens to critically examine regulative frameworks while fostering their trust. These various declarations of ethical principles are essential for guiding adequate and adaptable human-centred AI development with clear accountable parties.

The focus points and priorities of each of the ethical frameworks subtly change and fits in the general agenda of each organization. The IEEE global initiative for AI sets a

technical focus, mainly on transparency, accountability, and fairness of AI systems. The EU Guidelines on Ethics in AI put regulation in focus to ensure alignment with the broader EU's regulatory framework, including the subjects of privacy, non-discrimination, and harm. The focus of the OECD AI Principles is on governmental guidance toward developing AI policies and strategies. G20 AI Principles have an economic focus point and prioritize the potential growth, innovation, and international cooperation regarding AI development. UNESCO's Recommendation on Ethics of AI primarily focuses on human rights.

Nonetheless, these ethical guidelines are insufficient as a stand-alone to ensure adherence in practice. Effective enforcement strategies and the adjustment of the regulation are the next necessary step. Numerous national and international efforts and initiatives have been made the last years to establish different regulative strategies that are based on the proposed ethical guidelines. Particularly noticeable is the European Union with various efforts and regulation frameworks that encapsulate and promote the above-mentioned values.

### 2.2.2 European Union's AI Regulation

The European Union has been a significant player in the leading force of AI regulation and has made great strides with its latest **AI Act** which is the first comprehensive AI law for regulation of AI in the world<sup>16</sup>. With the regulation, the EU wishes to guarantee optimal conditions for the development and employment of AI systems and the compounded innovation. The lawmakers in the European Union have furthered tackled the regulation of AI with other instruments, e.g. the **Artificial Intelligence Liability Directive (AILD)** and the already established **General Data Protection Regulation (GDPR)**. There is an interplay between these three European tools, which to some extent do not fully align with each other but together promote and enforce the European agenda on its member states. The AILD intends to define compensation for damage caused by AI systems, either intentionally or negligently, that can be applied to providers and users of AI systems. GDPR is a regulation that covers legal issues concerning personal data of individuals within the EU. It aims to protect the privacy of individuals, limit the power of business over personal data, and ensure that the data is handled in a safely manner. GDPR in the context of AI, puts focus on the fundamental right to privacy, also in the era of big data and multi-layered systems that consume large amounts of data [92]. The AI Act is the most recent and significant legislation document worldwide that aims to protect and promote ethical values from crossing ethical guidelines.

---

<sup>16</sup>[https://www.europarl.europa.eu/pdfs/news/expert/2023/6/story/20230601ST093804/20230601ST093804\\_en.pdf](https://www.europarl.europa.eu/pdfs/news/expert/2023/6/story/20230601ST093804/20230601ST093804_en.pdf), accessed on: 06.08.2024

## EU AI Act [52]

The AI Act is the primary concrete legal framework for AI systems, as of 2024. The European Commission proposed the first EU regulatory framework for AI in April 2021<sup>17</sup>. Since then, further extensive negotiations have taken place that have led to an accepted version in December 2023 that was endorsed in March 2024 and finally entered into force in August 2024<sup>18,19</sup>. The EU AI Act promotes the digital strategy of the EU that sees great importance in ensuring human oversight of AI technologies and in a uniform protection of public interest and individuals' rights, which fosters the value of human rights. The EU AI Act's scope of application includes the *providers, users, and any other relevant party* for AI systems that are *located or used in the EU market*, which emphasizes the values of responsibility, accountability, and liability. In other words, also the AI Act still covers the systems that process non-EU data if placed within the EU market. This principle correlates with the product regulation approach that treats AI systems as European products, hence expanding the regulatory reach<sup>20</sup>. Furthermore, the AI Act acknowledges that some national initiatives have already been created and implemented on the internal market and hence aims to promote consistency among the different member states by determining a high-level framework that prevents significant divergences within the EU.

The EU AI Act takes a risk-based approach, namely AI systems can be used in different applications according to the analyzed risk they pose to users, which promotes the value of non-maleficence, safety, and security. For this, there are four introduced risk levels demonstrated in 2.3. They are concretely defined as following:

- **Unacceptable Risk Article 5:** AI systems that pose this type of risk are *completely prohibited* in the EU due to their detrimental nature of hurting human rights to an unacceptable degree. The exact banned systems are limited to a clearly defined set. Social scoring and systems that deploy subliminal components are examples for such systems.
- **High Risk Article 6:** AI systems that pose this type of risk are *partially permitted* insofar of compliance with the defined national and Union law. To determine the severity and likelihood of the risk, a *third-party conformity assessment* is provided. These systems have the potential to harm individuals' health, safety, or fundamental rights and are mostly used in critical infrastructure. The exact set of systems is clearly defined in the act. Among such systems are medical devices or tools used by the police.

<sup>17</sup><https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>, accessed on: 05.08.2024

<sup>18</sup><https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law>, accessed on: 05.08.2024

<sup>19</sup><https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>, accessed on: 05.08.2024

<sup>20</sup><https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law>, accessed on: 05.08.2024

- **Limited Risk (Article 50):** AI Systems that require *transparency obligations* to foster their trustworthiness. They have the potential to manipulate individuals, therefore the content has to be labeled as artificially created or manipulated. Such a disclosure has to be clear and distinguishable. Chat bots and “deep fakes” that have the potential to manipulate individuals belong for instance to this category.
- **Minimal Risk (preamble point 53):** AI systems that are *permitted without any additional legal restrictions*. They do not significantly influence the decision-making process and outcomes (automated or human-based) and cannot cause a substantial harm. Their scope of performance is tailored down to a narrow task that has no potential to pose a significant risk. In addition, these AI systems are intended to merely enhance a completed human-based task, and not be a stand-alone task.

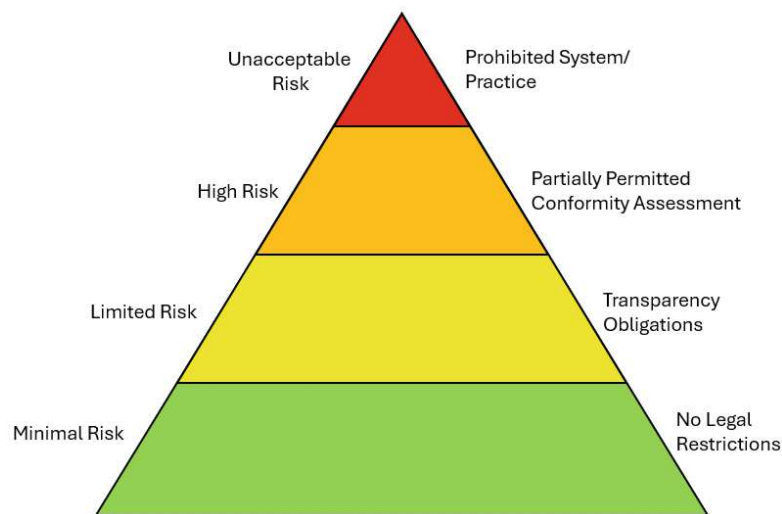


Figure 2.3: The four risk levels and the legal obligations declared by the EU AI Act

Another consideration of the act are the general-purpose AI models, namely models with large datasets that can perform a wide range of tasks and are not task-dependent, *in particular, large generative AI models, capable of generating text, images, and other content* (preamble point 105). These models have the potential to be further modified or fine-tuned. Thus, they can serve as the base for newer versions (also called downstream) that can be misused and become models with inherited risks, which regulation treats as systematic risks. The providers of these models have to perform model evaluation, mitigate systematic risks, document, report, and ensure a sufficient level of cybersecurity protection (Article 55). To determine whether a general-purpose AI model has systematic risks, there are several criteria, that are defined in Annex XIII, that the commission should consider.

The EU AI Act’s rigorous framework has the potential to shape the legislation landscape also beyond the EU scope. This piece of legislation is unprecedented and thus offers new

regulative benchmarks. Given the constantly evolving AI risk landscape, it is crucial to closely monitor the AI Act and ensure its relevance to handle these dynamic changes. The implementation of the AI Act is a key factor for its efficacy as well as understanding the evolving risk landscape.

## 2.3 Risks of Generative AI

Intelligent systems are the source of technological, economic, and societal growth while also providing the ground for many new feasible risks. The AI risk landscape is a dynamic field that has evolved dramatically throughout the last years and is compounded out of intentional and unintentional risks. Malicious actors have learned over the years how to exploit AI systems for various purposes, including financial crimes, biological weapons, and terrorism [58]. Naive employment of the technology could lead to known or unknown rooted bias to largely influence the outputs and causing as well an unintentional harm.

The debate about the potential risks of intelligent systems has been around for decades. In the early years, it was mostly a futuristic discussion filled with speculations that could be based solely on the existing expert systems, i.e. logic- and knowledge-based methods that use encoded knowledge for reasoning and modeling to solve a defined scope of tasks [149]. The discussion about the potential AI risks has reached a new phase since the technology surpassed the performance of a human-like level and the subsequent rising initiatives for establishing ethical guidelines and regulations. Leading concerns have evolved over time, driven by the massive transformation of the technological state-of-the-art along with the increasing geopolitical instability. A prominent risk in the current global geopolitical landscape is the potential misuse of Generative AI in the context of terrorism and warfare and how it can be exploited for radicalizing and polarizing.

The earlier risks included mostly the bias rooted in AI systems whose outputs are based on the training data they consume and the algorithms at their core. Warnings included results of an unchangeable status-quo that prolongs injustices dominating the world, referred to by Cathy O’Neil as “weapons of destruction” [132]. In the earlier stages of AI systems, the risk of poor standards for the quality and potential bias in the training data was slowly unfolding, as more cases worldwide hinted at discriminating decisions stemming from AI systems, in times with no disclosure for artificial intelligence involvement. An example of a famous case taking place in 2015 involved the recruitment scandal of Amazon’s “sexist AI” tool that automatically rejected submitted CVs of female candidates, based on the algorithm and the provided training data with the low representation for women in the company [1]. Another increasing concern of AI systems that need large amounts of training data and computational power is the high carbon footprint, as Bannour et al. [14] address. AI has been rapidly integrated into social-related systems with big responsibilities, e.g. organ implementation queue, scholarship distribution, acceptance to education institutions, etc., leaving more impacts on individuals at large. In addition, there has been a significant rise in surveillance

systems with the extreme example of social scoring [52].

Meanwhile, together with the latest developments of LLMs and Generative AI models, more concerns have been raised. The current historical-political events and conflict zones have also been of interest to academia, researching the ongoing and potential misuse of Generative AI technologies in this context. Ferrara's comprehensive work [54] from 2024 covers the nefarious applications of Generative AI and LLMs and their abuse potential. He presented a mind map with several categories of abuse; misinformation and disinformation, malicious content generation, bias amplification and discrimination, data privacy attacks, automated cyber attacks, identity theft and social engineering, deepfakes and multimedia manipulation, and financial fraud. Under the category of "Information Manipulation", one refers to a twisted information ecosystem that promotes deceptive content, such as misinformation and fake news, while exploiting the different advanced tools that provide an immersive environment. Chen [32] found that prior to 2012, there were no publications on the Scopus database addressing the intersection of social media and misinformation. However, in subsequent years, there has seen a significant surge in research and publications on this topic.

Ferrara identifies also "societal, socio-technical, and infrastructural damage" as a critical category. It can lead to the most catastrophic consequences impacting broadly communities, societal structures, and critical infrastructure, as well as harming democratic processes and social harmony. The intent of propaganda is also addressed, mostly in the context of using falsified facts or manipulating sentiments to promote political, ideological, or commercial interests. Ferrara mentions the risk of having automatically generated propaganda campaigns that governments or organizations utilize to target specific groups in a personalized manner and produce diverse material at a high scale. Nevertheless, terror organizations as potential malicious parties are not mentioned in this paper, leaving open questions about the use of generated AI propaganda for terrorist activity among other possible utilities [54].

Lakomy identified in his paper from 2023 [98] this gap of research about terrorist exploitation of AI while emphasizing the high volume of research about the use of AI for counter-terrorism. Trying to fill this gap, he carried out an exploratory study in which AI technologies (LLMs and image generators) of open-access were tested on terrorist-related tasks, such as producing and distributing terrorist propaganda, or providing access to terrorist content and know-how. The image generation study investigated DALL-E 2, Dream.ai, and Bing Image Generator and focused on the generation of visual propaganda. He found that it is possible to reach visuals that aesthetically match violent extremist groups, including known symbols and logos, and reach a hyper realistic image quality of militants that can be re-used for terrorist campaigns. Lakomy also recognized the integration of anti-terrorism regulation in the open-access platforms that got tested, as attempts of prompts including clearly terrorist terms or the copying of terrorist aesthetics were blocked. The generation of terrorist combat images that involve humans were also

blocked as well as any explicit death scene. These results are first of a kind and only relevant to the above-mentioned open-access platforms in 2023 without going any deeper into the different types of propaganda.

Puczyńska et al. also published a related work [143] in which they investigated the use of LLMs in jihadist terrorism and crimes. They see potential in exploiting the technology to recruit terrorists by radicalizing and polarizing with personalized information bubbles, applying social engineering for scamming, creating, and distributing disinformation. They put focus on textual content that could be exploited for those purposes, including translation tasks to reach large and diverse audiences. Puczyńska et al. briefly mention the use of graphics for terrorist use and propaganda and warn about the potential of AI to contribute to these efforts. A concrete example given in their work for this case is generating manipulated graphics of suicide terrorists with a happy face, hiding any traces of the physical suffering that aims to motivate potential individuals to follow this path for the terrorist cause. Also Esmailzadeh [50] briefly warns of the risk of deepfakes for propaganda and manipulation and adds that such fake videos could help portray the enemies in a negative light.

In a similar context, Makhortykh et al. [111] published a paper that touches on the future of mass atrocities' collective memory and the relevant risks. Similarly to Ferrara, they also touch on the importance of being able to differentiate between human- and AI-generated content as well as the risk of promoting false information. All of these studies warned from terrorist misuse potential of generative AI technologies, in particular of visual deepfakes. However, none of these works explained the fundamentals of propaganda or how this manipulative information manages to reach the mass crowds and how they react emotionally on graphic violent scenes. It is thus of interest to understand how these techno-psychological mechanisms play to the hands of terror organizations and their purposes.

Further outstanding risks that Ferrara addresses include the generation of deep fake porn, fabrication of historical facts, and rewriting of history through authoritarian regimes and complete control over the flow of information, automated and coordinated content on multiple social media platforms, and misinformation for populist interests and election campaign distortion. Ferrara also leaves technically concrete recommendations for the mitigation of GenAI abuse. These include:

- **Proof of identity** to ensure traceability of the content creation or other AI-powered actions
- **Authentication protocols** to confirm the identity of the content creator or responsible party for AI-powered actions
- **Audience disclaimers** to inform the viewer about the AI origin of the content and to promote transparency that enhances the critical consumption of content

- **Content labeling** to distinguish the AI-generated content
- **Source verification and provenance** to keep track of the course of ownership to maintain integrity and trustworthiness
- **Digital watermarking** to identify AI-generated content and add a layer of security, integrity, and traceability

These recommendations are relevant only in the case of cooperating actors who would be willing to implement these technical measurements. An additional recommendation is a risk-benefit analysis that examines the short-term and long-term influences of the potential harms. Ferrara focuses on the importance of identifying AI-generated content to protect societal values and norms.

There is also little work about the risks of perpetuating violence by Generative AI systems. Google researchers Srinivasan et al. [72] published in February 2024 a recent study concerning this issue. It addresses the capability of Text-to-Image models to produce unintentionally amplified harmful visuals, while also giving a clear definition to the term “harm amplification”. The need for developing socio-technical safety requirements is addressed in this work alongside demonstrating generated figures that are notably more harmful than the intended input text.

Also, sexual violence-related risks are relevant to perpetuating violence and belong to cyber violence and are practices seen in war zones. Hunter [81] warns that the technological ability to generate AI porn sets a high risk, especially for women. Additionally, in their paper [19] from August 2023, Bird et al. address the risks of Generative Text-to-Image models, naming 22 distinct risk types, including sexualized imagery, including child sexual abuse imagery, violent or taboo content. The latter focuses mostly on violent content against minority groups, or politically incorrect images. Bird et al. address as well the risk of misinformation and disinformation, including individual, social, and community harm.

The discussions about the risks of generative AI also found their way out of the academics and received a main focus by impactful global reports. The internationally recognized World Economic Forum’s (WEF) global risks report for 2024 [58] conducted several surveys, including varied stakeholders; civil society, international organizations, academia, government, and private sector. Among the noticeable results, AI-generated misinformation and disinformation are in the second place of the current main risks, making 53% of the votes and societal and/or political polarization in third place with 46% of the votes. Moreover, misinformation and disinformation were identified to be the risk with the highest severity in the short term and polarization on the third place. WEF’s report acknowledged the interconnections and their density between the different risks. Misinformation and disinformation are highly connected to societal polarization, and further connections were found to intrastate violence as well as censorship and surveillance,

and erosion of human rights. On the secondary connection level to misinformation and disinformation, one can find terrorist attacks and interstate armed conflict.

From the above, we learn that all forms of information manipulation can lead to harmful consequences for society and democracy. Several malicious parties have an interest in exploiting this risk, including authoritarian regimes, political actors, especially during elections, and terror organizations. The link of manipulated information has been linked to radicalization and polarization, mostly for terrorist causes. The perpetuation of graphic violence through Generative AI has lately received more focus in research. Moreover, there is new evidence showing that the current state-of-the-art has the tendency to create more violent outputs than the intended input and is in the state to create sexual pornographic content. The full scope of risks of generated graphics in radicalization and polarization processes and the further psychological implications to society, in particular by terror organizations, remains incomplete. Our research aims to address these gaps and reach a holistic overview that includes the entangled societal and technological processes.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Methodology

We used the method mix [39] to answer the two parts of our research question. We first investigate the risks stemming from the exposure to visual violence and the following radicalization pathway, along with all of its components (RQ1). We then utilize our findings and search for the concerning general and technological gaps in the recent legal framework of EU AI Act (RQ2), and propose explicit technology-specific solutions to bridge the found gaps and mitigate the identified risks (R3).

The literature suggests to combine both quantitative and qualitative research techniques to provide a comprehensive understanding of the topic of research through providing both a broad and a nuanced overview. The method mix allows us to obtain on the one hand a broad overview of the subject of visual violence and its role in radicalization while identifying societal patterns and on the other hand gives us the ability to delve deeper into the resolution and nuances of the specific legal document.

First, we conducted a **systematic literature overview** using **PRISMA** (Preferred Reporting Items for Systematic reviews and Meta-Analyses), that Liberati et al. [104] explain.

The database for this work is constructed through applying our defined search term on two main digital libraries; IEEE and Web of Science. The initial exclusion criterion is if the paper is not available. The initial inclusion criteria are:

1. Papers are in the English language
2. Papers appear in a journal or a conference proceeding

The applied **search-term**: *(radicalization OR terrorism) AND (AI OR social media OR algorithms OR algorithmic)*.

### 3. METHODOLOGY

The researched categories: *Title OR Abstract OR Keywords*.

Initial found articles from:

1. **IEEE:** 294 articles for searching in abstract, 31 Author keywords, 15 document title (in total 340 articles)
2. **Web of Science:** 1355 articles for topic (defined as title, abstract, keyword plus, and author keywords)
3. **Scopus:** 1626 articles for Article title, Abstract, Keywords together

In total, there are 3275 papers. Without duplicates, there are 2606 articles in the initial database. We proceed with further exclusion of records based on the titles and exclude and are left with 267 papers.

In the next step, we exclude further records based on the abstract and are left with 109 papers. In the final step, we read the full text of each of the collected records and included in the final database only the relevant articles, in total 73 papers. We demonstrate in 3.1 our steps toward building a working database.

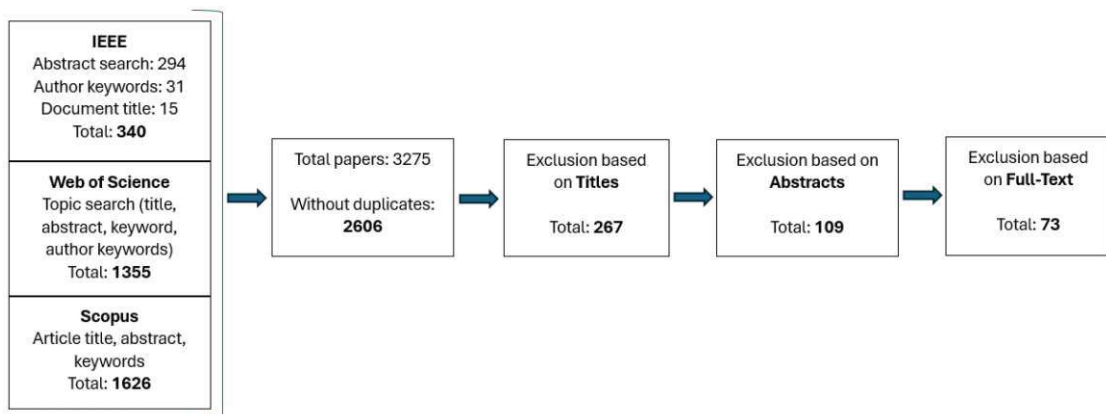


Figure 3.1: The PRISMA screening steps toward a database

For some of the topics, we decided to proceed with further research. This helps us understand specific topics that did not directly come up in the results of the search term but the papers in the corpus briefly mentioned them. We manually searched these topics on Scopus and included also books.

Using the analyzed corpus, we identify the most relevant information to radicalization, polarization, and psychological harm. We then integrate our acquired knowledge

---

with the outlined risks in the initial literature review presented in the background 2.3. This approach allows us to draw inferences on the setting of these risks and deepen our understanding of the elements and factors that exacerbate them.

We define each risk with our own proposed framework, which includes the assigned risk level, the responsible key actors, whether the harm is direct or indirect by these actors, the contributing context, and the harmed target group.

Second, we conducted a **legal document analysis** following Bowen's work [22]. This qualitative research method is the second complementary component of the method mix. This method is a systematic procedure that aims to review and evaluate printed or electronic documents. It usually involves the iterative process of skimming, reading, and interpretation.

Our chosen data source is the latest legal framework, the EU AI Act, which is a given text without our intervention. Our reference point is the current latest version of the EU AI Act from June 2024 [52]. There is typically a prior literature review, which we have conducted with our first methodological part with the PRISMA framework. We then incorporate the information we gained from our identified risks (RQ1) into the document analysis.

The analytic procedure encompasses finding, assessing, and synthesizing data in the document. Relevant and meaningful data includes excerpts, quotations, or complete passages. There are a few categories that we wish to cover in our analysis:

1. Actors covered by the EU AI Act that are relevant to the risk actors
2. Technological components covered by the EU AI Act that are relevant to the risks. These include the five points raised in the background as primary challenged of generative AI systems, namely bias, misinformation and disinformation, privacy, transparency, and explainability. Based on the mitigation of these points, we can base our findings concerning the technical requirements.
3. The mitigation of risks, based on the risk-based approach and its relevance to our identified risks

With this method, we draw a line between the covered elements in the EU AI Act and their relevance to our identified risks, while identifying gaps or open points that are relevant to our risk mitigation.

In this analysis, we focus on both the defined technology-specific standards in the regulation, concerning the key points of the brought technological challenges of generative AI systems, and the relevance of the risk level definitions to our identified risks. We prefer

### 3. METHODOLOGY

---

to exclude the efficiency of the regulation itself and leave it to future work. This way we follow a selective approach that focuses on the three categories mentioned above with regard to the concerned risks. We identify technology-specific and general gaps in the EU AI Act.

Our final step involves aggregating the acquired technical knowledge from the literature review and the PRISMA corpus, and provide technical solutions to the identified risks and open technological-specific standards, which we deem essential to address. We focus on providing technology-specific recommendations by drawing on our understanding of the inherent challenges in generative AI systems, as well as the possible technological solutions addressed in our systematic literature review.

Through this iterative analysis process, we aim to produce empirical knowledge that highlights the current risk landscape of generative AI systems in the context of warfare and terrorism, the technological and non-technological gaps in current regulation with regard to the identified risks, and necessary technical requirements to enhance safety and resilience against extremist misuse of generative AI.

CHAPTER **4**

# Analysis

In this chapter, we analyze the knowledge base of the corpus created with the PRISMA framework. We observe our corpus as a socio-technical system with which we systematically understand how the different social and technological factors involved mutually affect radicalization, polarization, and psychological processes. The people component serves as the first layer in which we learn that everyone is prone to radicalization and therefore it should be rather observed as a process. The radicalization models consist of several steps, including mostly the enhancement of the initial micro-level discontent, a shift in mind toward a twisted interpretation including an external factor to blame, the comfort found in the group's belonging, and its affirmation. There is a significant role in enhancing the 'us vs. them' feeling in the radicalization process. This feeling is also key in political polarization processes. The internet infrastructure and with it the diverse social media platforms belong to the technology component. We find that algorithms that initially aimed to personalize cyberspace can be an important contributor to radicalization. The feedback loops, filter bubbles, and echo chambers are all components of algorithmic radicalization through which individuals are exposed to one-sided content, and the effect of this exposure can be only enhanced by the concepts of anonymity and decentralization that enable the overload of vast amounts of unreliable yet convincing information, but could also preserve privacy. Algorithmic detection techniques of extremism content and violence outputs can serve as valuable safeguards when integrated into generative AI systems, though they present technical challenges. We follow the roots of terrorist misuse of these societal and internet tendencies and learn the structured manner of terrorist organizations to target prone individuals through manipulation, mainly for recruitment purposes. The last layer of our socio-technical systems integrates all of these elements and demonstrates how they can be navigated for propaganda and polarization and what psychological impact they can have. The employment of graphic violence from live-streamed atrocities as the means of propaganda itself is a recent phenomenon that has become more evident in current warfare and terrorist activities. Unstable times are often followed

by political and social instability that polarization emphasizes. The contrary tendency and a possible solution for the division in society is the social resilience that brings the hearts closer together, emphasizes similarities, and largely contributes to the continuous functionality of society in hard times. The present graphic violence triggers various psychological phenomena that in the extreme case leads to the glorification of violence, encouraging individuals to act similarly eventually having a large impact on society. The exposure to extreme graphic violence increases also national-level mental health problems, mostly including short- and long-term post-trauma that have been mainly studied after the 9/11 attacks and have newly received focus, especially after the graphic images and videos from the 7th of October attack. With this comprehensive understanding, we lay the theoretical foundation for our risk analysis and provided technical solutions.

## 4.1 Socio-Technical Systems

The corpus of the selected papers using the PRISMA framework unveiled the multifaceted ways in which generative AI can contribute to radicalization. The aim of our analysis is to **construct the pathway of radicalization stemming from the exposure to visual violence, that the manipulation by the recent Generative AI technology often enhance**, the interplay of social media platforms and terrorist actors, and the psychological aspects and harms of this process, as 4.1 demonstrates. By constructing this pathway and analyzing the components that contribute to its “input” and “output”, we can comprehensively assess the full scope of the risk landscape that terrorist misuse of Generative AI causes.



Figure 4.1: Radicalization through Exposure to Visual Violence Pathway

Therefore, this analysis, we will adopt a socio-technical lens, that Eriksson Krutrök and Lindgren [49] apply in the context of misinformation and Risius et al. [151] in the context of online extremism. We argue that the a socio-technical perspective is well-suited for examining Generative AI-powered radicalization. This technology enables the rapid production of vast amounts of new and manipulated information, that malicious parties can easily misuse. Coupled with the online infrastructure, which is essential for widespread information dissemination, the mutual dependencies between the social and the technical components become evident.

Bostrom and Heinen [21] introduced in 1977 a socio-technical perspective combines social elements and technical elements with mutual influences and dependencies. The social components are divided into **people**, which are the individuals and the social processes, and a **structure**, which stands for the groups and organizations. The technical

components are divided into **technology**, which is the technological and physical systems, and the **target tasks** of the organized structure [25]. Risius et al. [151] applied the socio-technical system of Bostrom and Heinen to the context of terrorism and extremism. We take a similar approach and demonstrate in our adjusted graph 4.2 the connections of the different socio-technical components in the context of online radicalization as a result of exposure to visual violence that we will further explore throughout our analysis.

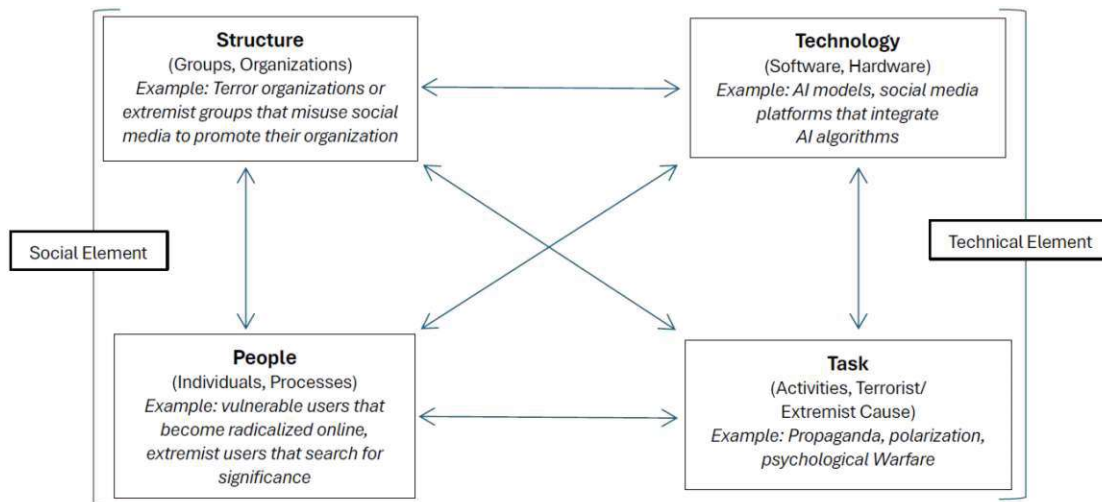


Figure 4.2: The interacting socio-technical components of online extremism adjusted from [21]

Furthermore, we will apply Bostrom and Heinen’s classification to the radicalization pathway as follows:

- **People:** individuals who are prone to radicalization, radicalization models, and processes. In this category, we will observe the role of individualistic perspective within the radicalization pathway.
- **Technology:** algorithmic radicalization and extremism detection that dominate cyberspace and social media platforms.
- **Structure:** terrorist misuse of social media platforms and the well-organized coordination of terrorist activities that online infrastructure in the media environment enable. In this category, we will also discuss the evolvement of the systematic use of extremely violent visual content for propaganda purposes of terror organizations.
- **Tasks:** terrorist activities aiming at radicalizing, polarizing, and finally causing harm to society. We will focus on the psychological harm and further effects.

Individuals, technology, and organizational structure have a direct influence on the “input”, while tasks align with the “output” of our radicalization pathway. We discussed the

significant capabilities of Generative AI models and their visual outputs in the background 2.1 where we have also seen that there is little work with a narrow understanding to the terrorist use of the text-to-image and text-to-video Generative AI models. Therefore, we will analyze the use of infrastructure that social media provides for the varied terrorist causes through a the socio-technical lens to address this gap. We will finally incorporate the additional and recent layer of generative AI. This systematic analysis would help us draw the concerning evolving risk landscape.

### 4.2 PEOPLE: Individuals and Radicalization Processes

The first social component, the people, serves as the first layer for the radicalization pathway. From our working corpus, 11 papers were relevant to the people component in our socio-technical system. We found many intersections between the different sources for explaining radicalization theory. Therefore, for this part, we proceeded with further research that follows and combines the initial literature that served the papers in the corpus. The scenario of an individual who consumes content online and becomes radicalized and extremist is more complex than it seems at first sight. The definition of radicalization has to be fully understood as well the attributes of the individuals who are susceptible to radicalization and the factors or processes that lure them into engaging with extremism. Another frequent matter in this context is the question of engagement, what is the threshold that drives individuals into action and what are the risks of extremist that do not engage in physical violence but hold radicalized opinions.

#### 4.2.1 The Semantics of Radicalization

The concept of radicalization remains ambiguous when it comes to a clear deterministic definition to be internationally used and it has developed over time as different historical events occurred, as Neumann [129] explains. In this work, we mainly follow Borum's definition [20] that he introduced as the development process of extremism of *ideas or behavior*. This definition is quite abstract and we can get more specifications by looking at further definitions. Constanza [38] defined the process of radicalization as the progression over time that enlightens how an individual or a group shifts towards radical beliefs in a high-paced society. Mccauley and Moshalenko's work from 2008 [114] addressed that common radicalization nowadays mostly comprises non-state parties challenging the state authority and they understand it as "change in beliefs, feelings, and behaviors in directions that increasingly justify inter-group violence and demand sacrifice in defense of the group". We see that according to the different definitions, not in all cases the extremist ideology turns into extremist behavior, hinting at the existence of different types of radicalization as well as the different approaches toward it. The literature often refer to it as cognitive radicalization and behavioral radicalization [139] [186]. As for violent extremism, there is no global consensual definition between academics and policymakers [69]. The United Nations [27] defined violent extremist as "someone who promotes, supports, facilitates or commits acts of violence to achieve ideological, religious, political goals or social

change”. Kruglanski et al. [94] and Al-Saggaf [5] suggested that radicalization exists on a continuum in which people engage mentally with the ideas and support the accompanying violence to the point of physically exercising violence in different grades. Following the papers of Borum [20] and Leistedt’s[100], the combination of various factors, the “why”, which drive the whole process forward by different components, that together construct the means, the “how”, motivate the radicalization process. The latter became the main focus of radicalization studies, further explored in 4.2.2. This shift happened since Horgan’s article from 2008 [76] following the idea to study the process and not the common grounds.

Radicalization is not a 21st century phenomenon, there are multiple examples, both from democratic and non-democratic societies illustrating its implications throughout history. The roots of “individual terror” are in the late 19th century, bringing along a wave of a new kind of violence called “political terror” that was different than the ancient “political assassination” by having complex ideological motives. The most famous individual act of terrorism is the assassination of Archduke Franz Ferdinand in 1914 by Gavrilo Princip in Sarajevo which led to the start of WWI [85]. The rise of Nazism and the applied propaganda methods at the time were also at the center of attention in the social sciences being an exceptional example of a large-scale group radicalization. This example also demonstrates the great importance of preventing a repeated situation in which an entire population was brainwashed and radicalized, both in mind but also behavior, from happening again.

We explored various definitions of radicalization, focusing on the transition to extremist beliefs, violent behavior, and even terrorism. While our primary interest is in terrorist activity, we also consider non-terrorist forms of extremism to achieve a comprehensive understanding. To gain a holistic perspective, it is essential to deep dive into the attributes of the individuals and groups being prone to radicalization as well as the abstract components of radicalization models.

#### 4.2.2 Target Groups of Radicalization

The main focus when studying the target groups of radicalization is those who engage in terror acts. In earlier years, there was a tendency to refer individual terrorist activity to psychological or psychiatric anomaly [20]. However, the existing evidence shows the contrary, namely, most people who fall for terrorist activity are rather rational [160] and psychopathology speaking do not deviate from the average individuals in society [45]. Observing ISIS’ case, the US recruits are from a varied demographic background, considering their socioeconomic and ethnic origins [117]. Nevertheless, young people are the most vulnerable group to engage in radicalization and violent extremism, even while many of them attend formal education [155]. Further research of the root causes included different theories and studies, as Kruglanski et al. [93] overviewed. For instance, studies have shown that lack of formal education or poverty is not a necessary sole factor for radicalization to terrorism. In addition, the theory that claims that those who are frustrated become aggressive enough towards others to engage in terrorism is

also no longer statistically admissible as a stand-alone, but it rather serves only as one of the necessary pre-conditions. The different sought directions of the root causes that characterize those who become radical have not built up strong empirical evidence. These directions are called **contributing factors**; under certain combined conditions and some circumstances they might lead the individual to choose the path of engaging in terrorism.

This change in the observation when studying radicalization hints at the large number of people who are prone to radicalization. Doosje et al. [45] even claim that “there is a terrorist hidden in everyone”. They further argue that there is an additional layer, referred to as **roots**, that also play a significant role in being prone to radicalization. The roots consist of three layers:

- **Micro Level** refers to the individual factors, mostly include the search for significance and feelings of uncertainty that an external group can be easily provide.
- **Meso Level** refers to the support and validation given by the close environment and the feelings of identification to the group as an influencing factor. This is especially the case when the feelings of injustice are enhanced among the group members.
- **Macro Level:** refers to a larger influence by societal tendencies or governments that often includes the globalization effect as a threat to the group identity.

Recognizing the factors and roots alone is insufficient to fully explain what brings individuals to engage with radicalization and terrorism. This acknowledgement rotated the approach to study the topic and sent it in a new direction in which radicalization studies should be learnt with a focus on learning the process through which individuals become involved in terrorism.

### 4.2.3 Radicalization Models

The process of radicalization is empirically invalidated, however, most approaches agree on its multi-step sequencing characteristic. The process of radicalization is not unified for everyone, hence, the same pre-conditions will lead to different radicalization mechanisms for different individuals [115]. Sageman [160] adds that external factors are an influential component in the radicalization process, but states that at the same time, those factors concern millions around the world who do not engage in terrorism. Moreover, there is an almost inverse path to radicalization, namely, some join radical groups to express their tendency to violence and even rationalize it, as explained by McCauley and Moskalenko [115]. The 3N radicalization model and Sageman’s radicalization model are excellent examples for a three- and four- steps models that often serve as a reference in the literature.

### The 3N Radicalization Model

The 3N radicalization model of Webber and Kruglanski [185] stands at the core of many radicalization studies and many researches have examined and accepted it. They proposed a radicalization model of individuals that is composed of a trilogy of psychological forces that are empirically supported:

1. **The needs or motivation.** These needs or motivations are mostly religious or political. The different offered individual motives, e.g. honor, humiliation, injustice, etc, are varied, but at the same time, they all hold in common the wish for significance and finding self-worth. Feelings that are group-based refer to the group identity and the related feeling of belonging. This process begins after the occurrence of some triggering event that involves a significance loss (the feeling of being insignificant), potential loss of significance (being motivated by the ability to prevent the potential loss), and significance gain (the accompanying hero status earned in the group). The end result is an individual urge to act violently in the name of a collective and it corresponds with feelings of insignificance infusing the rising need for closure.
2. **Ideological narratives.** This component thrives within the cultural context of the individual. Often, the societal shared values and norms of the individual restrict the way of thinking and the ability to find new solutions to dealing with negative feelings and not following the already existing pathways. When the group and its promoted values are rather violent or radical, actions that are out of the moral norms of parallel social units can be permissible and socially accepted by using means of delegitimizing the “others” using the strategy of dehumanization.
3. **Group dynamics and social pressure.** Ideology needs space for expression and consensual validation, otherwise, it cannot thrive and is doomed to die out. sharing radical thoughts and finding comfort in the confirmation from peers is the essence of group dynamics. The personal identity of such individuals transforms into their group identity, giving even the feeling of a second family and merging their individual needs with those of the group.

### Sageman’s Radicalization Model

Sageman [160] is one of the initial experts to have come up with a radicalization model that follows the bottom-up approach and sees the important role of social media and the internet in radicalization to violent extremism. His model has been used as well for the base of many analyses and is there for an important milestone in radicalization studies. Sageman focuses in this model on Islamic radicalization and characterizes the process as a four-staged:

1. **The feeling of moral outrage.** This stage includes an emotional reaction to a negative event or course of events, on the individual level or on the global level. The perception of the two levels might merge together.
2. **Single interpretation of the moral outrage as a war.** The “othering” mentality gets a rise, leading to strong “us” vs. “them” feelings.
3. **Continuous experiences of discrimination and bias.** Often combined with financial problems and unemployment. These circumstances lead individuals to perceive themselves as a victim of multiple factors (e.g. social, economic, political).
4. **Establishment of social network.** These factors create a turmoil of frustration that brings together individuals with similar feelings and extremist ideologies. Especially thanks to the internet, the space to connect is easily and globally established which eases communication.

Both models share common components, including the initial individual discontent, the shift in worldview, the reinforcement of one’s own identity over that of others, and the great importance of group belonging in which these accumulated feelings are expressed. Both models emphasize the significance of “othering”, particularly through *dehumanization* of the other group. Roberts-Ingleson and McCann [152] argue that there is a link between misinformation as a tool for enhancing emotions in multiple contexts for different objectives and the process of radicalization. They claim it this phenomenon is notable to individuals who actively seek to engage with radicalization. On this note, we recognize the exploitation potential of Generative AI to amplify the key components of radicalization, serving as a social element within a socio-technical system.

### 4.3 TECHNOLOGY: Infrastructure and Radicalization

The first technological component, the technology or physical systems, plays a significant role not only in the creation of visual misinformation by Generative AI, covered in 2.1, but also in providing the infrastructure to effectively disseminate it. The latest technological developments, especially the Internet of Things (IoT) and cloud computing that support scalability, have led to the expansion and globalization of social interaction together with the easy distribution of extremist narratives. Generative AI technologies also highly depend on cloud services, as the training of the models largely runs on the infrastructure of external cloud services (GPU cloud), following the principle of “Infrastructure-as-a-service”. These cloud services address the needs of large AI systems to have robust computational resources and scalability [184].

The latest historical COVID-19 pandemic and the accompanying worldwide lockdowns gave an additional technological and social push and emphasized and even increased the importance of the cyberspace in communication, not only on an international level but also on a national one [135]. Social media has as well changed massively throughout

the years and is no longer the sheer simple concept of a platform for messaging and content exchange and the current state-of-the-art involves more and more smart features. With time the complexity of the main social media platforms has gotten even out of control, with millions of code lines per platform<sup>1</sup>, including complex databases, bots, and sophisticated algorithms [173]. Proctor [142] named this new environment as “cybernetic animism” in which non-human components dominate and the humans in this ecology cannot distinguish with whom they interact. These interconnected changes are social and technological which continuously and mutually influence each other. In our corpus, there are 14 relevant papers to the physical systems and they primarily focus on the built-in algorithms at the core of social media platforms. These algorithms, which often involve AI techniques, act as technological enablers and facilitators of radicalization [166]. These algorithms, responsible for determining the shown content based on profiling of the user, following the principles of “recommender systems”, have far-reaching social implications. Additionally, some papers have explored the parallel technical and social concepts of anonymity and decentralization, which form a socio-technical infrastructure that can also contribute to the radicalization of individuals online. Detection techniques of textual and visual content can be helpful safeguards in inhibiting violent visual outputs of generative AI systems but also pose a few technical challenges.

#### 4.3.1 Algorithmic Radicalization

Online and social media algorithms are built in a way that maximizes the satisfaction of the user. For this cause, different algorithms were designed to be aligned with the users’ wishes and many social media platforms follow the idea of recommendation systems, as Zhang et al. [190] introduced in their work. We get nowadays recommendations throughout our user experience for dozens of service components, from what to cook, to what content might be interesting for us or whom we might know and shall connect. Furthermore, the search has become personalized by keeping digital profiling based on the history of interaction of each user that is in some cases cross-platform and aims to find the most relevant results for each user. The increasing use of centralized identity providers also plays an important role in the personalization of the internet and enhances the profiling of digital identity in a centralized way [182] [156]. These algorithms have become an integral part of the ecology of the internet and in particular of social media platforms [173]. These algorithmic implementations started with rather financial motives and were quickly exploited by turning them in another direction. The developed algorithmic structure that stands at the core of the internet and social media platforms have led to different socio-technological phenomena

#### Feedback Loops

Recommendation systems largely support the continuous activity of a user on the online platform by suggesting further similar content. Also called “amplification loops”,

<sup>1</sup><https://informationisbeautiful.net/visualizations/million-lines-of-code/>,  
Accessed on: 17.05.2024

as Eriksson and Lindgren [49] introduced, feedback loops lead to a cyclic pattern of segmentation into topics that are based on the characteristics of the users. This technical implementation is activated by the user's behavior and accelerated by social media platforms who keep track on the digital identity and characteristics of their users. Feedback loops leads to the amplification of pre-existing beliefs and biases by reinforcing them through repeating exposure to the same content. Shin and Jitkajornwanich [164] illustrated the mechanism of a feedback loop 4.3. In their research, Hosseinmardi et

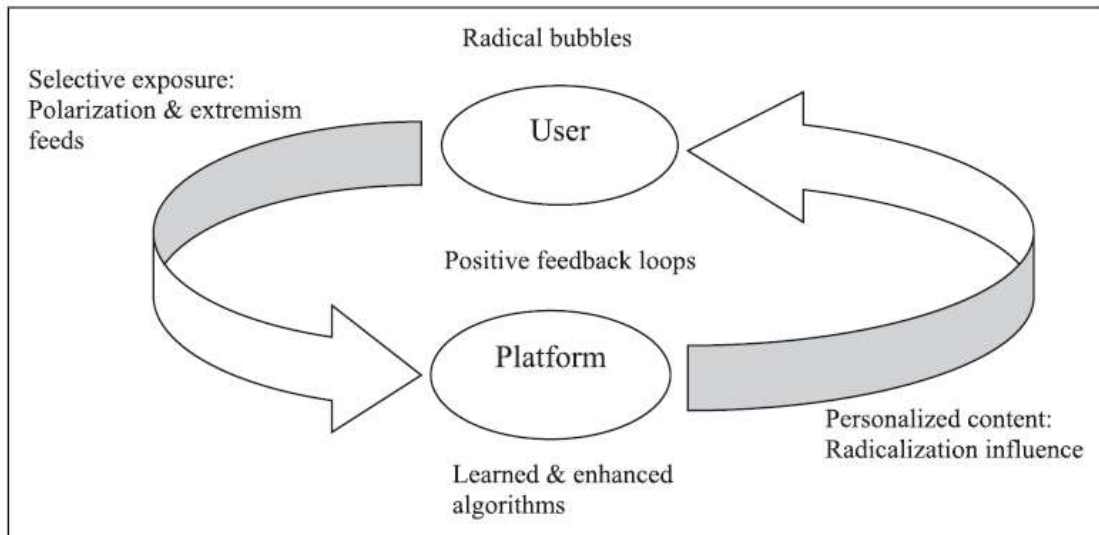


Figure 4.3: Shin and Jitkajornwanich's Mechanism of Feedback Loops [164]

al. [77] confirm that the recommendation mechanism tested on a streaming platform repeatedly direct users to more extreme and polarizing content and effectively radicalize them. Santos et al. [162] also found a link between recommendation algorithms and polarization by simulating network topologies.

### Filter Bubbles

Pariser [136] coined the term “Filter Bubble” which is the automated *algorithmic* process of selecting and showing the content that is the most relevant to the user based on profiling methods and databases. This means that the selected content is pre-filtered and creates bubbles of flowing content that is tailored down to common interests or characteristics of individuals. Feedback loops lay the technical base for filter bubbles. This also creates a misrepresentation of the “open and free knowledge world of the internet” whilst individuals are exposed solely to similar world views, cutting off the possibility of reaching varied information and perpetuating interaction with those who hold similar opinions. Miller and Record [121] also presented the layer of nontransparent filters as threatening back in 2013, because many are not aware of them. Those who do not know how these algorithms are designed and cannot determine or even underestimate

the degree of their isolated exposure. Filter bubbles are created based on the nature of the internet and social media platforms that enhance personalization and not because of malicious intention<sup>2</sup>.

### Echo Chambers

Filter bubbles endorse secluded virtual communities that limit access to a pluralistic environment. Jamieson and Cappella [86] coined the term “Echo Chamber” in 2008 which describes how the voices in those groups are internally driven and intensified, that Jong and L.A. Dückers [88] also referred to as “echo-effect” by. Sullivan [173] explains that in the environment that the echo chamber creates, the individuals actively choose to expose themselves to content that corresponds with their beliefs, and through that they obtain repeated confirmations that just stir up their worldviews while neglecting and even cancelling contradicting ones. This phenomenon is often accompanied by disregarding and attacking the traditional mainstream media and consequently contributes even more to the post-truth era [130].

Once again, a misrepresentation of the actual reality occurs, as Von Behr et al. [183] and [178] explained. There is an illusion on the surface of having a pluralistic and inclusive society and that the topics are handled as multi-faceted, while the actual picture is reversed and this only strengthens the feeling of having a consensus based on the perception of those individuals while cultivating the exclusion of any other external opinions. In Baugut and Neumann’s study [15], individuals believed they were not susceptible to media coverage influencing their ideas and opinions, while also expressing confidence in their own judgment. Nguyen [130] explains that an echo chamber is a social epistemic structure in which the voices of the “others” are intentionally excluded through manipulation of its reliability and even addresses the similarity of this mechanism to the one of a “cult indoctrination”. Simultaneously there is an internal growing and reinforced trust in the members of the community that creates a strong interdependence of the group members.

Gunton [69] covered the different approaches toward filter bubbles and echo chambers and he showed that their creation highly depends on the structure of the social media platform. It is therefore strongly believed that their dominance over the internet infrastructure provides the facilities to strengthen and reinforce believed narratives rather than directly leading to radicalization. Baumann et al. [17] modeled the polarization dynamics in the brain and found that for controversial issues, the reinforcement taking place through the echo chamber leads to radicalization and emphasizes the gaps in opinions to topics of consensus. Both terms function differently, however, can both be practiced in and influence the same groups [130].

<sup>2</sup>Nguyen’s article [130] deals with the social aspect of filter bubbles, namely epistemic bubbles, and gives an overview about its development in the offline world.

### 4.3.2 Socio-Technological Infrastructure

Different sociological structures of the offline world seem to have parallel technological components that are part of the internet infrastructure or social media. The interplay between these two realms is of interest when studying socio-technical systems. Anonymity and decentralization are both key concepts that are present in our corpus. They are prevalent both in digital environments and offline settings.

#### Anonymity

The opposite technological concept to filter bubble, which is all about personalization-oriented internet, is anonymity. It is a significant component of the internet and often serves widely used security and privacy-enhancing mechanisms [147]. This concept is implemented through encrypted messages that are transmitted over the internet or the ability to engage in conversations or one-sided forum comments without leaving public tracks on the identity of the publisher [173]. This ability increases additionally the online violent extremist radicalization, as it takes out barriers for some who are forbidden from engaging with extremism, for instance, groups of women from countries in which their rights are restricted, according to Von Behr [183].

Additional anonymity mechanisms for preserving privacy concerning visual attributes are widely researched, in particular for the domains of health care, social media, and surveillance. The blurring and pixelation of faces are mechanisms that are widely used as well as blacking out the facial area. The current AI and generative AI algorithms are already capable of inverting back the bits and come up with photos similar to the original [9]. Another problem of the blurring and pixelation techniques is the reduced photo quality that is often unwanted. The generative adversarial networks (GANs) brought new advancements also to the anonymization domain with data synthesizing techniques. Hukkelas et al. [79] presented a novel architecture with their *DeepPrivacy* conditional GAN model, whose main idea is removing the visual attributes that are sensitive information with their personally identifiable information (PII). These visual features are then exchanged with synthetic data, i.e. visual features of generated individuals, maintaining a high image quality, as seen in 4.4.



Figure 4.4: An example for the anonymization with the conditional GAN model

In this model, the generator observes the given background of the image and repeatedly generates a realistic set of anonymized faces in the correct pose based on key points

(the position of the ears, nose, eyes, and shoulders in the space of the image), while not having access to the original face in the input picture. The discriminator includes the background information as the conditional input as well as the pose information of the face and repeatedly evaluates the quality of the photo against a validation set [79].

Another connected phenomenon to anonymity is a lesser repercussion when sharing content, as users tend to perceive their online and physical identity distinctively [173]. Considering the post-truth era discussed earlier, the credibility of the source is no more relevant for trusting and accepting the information and the component of anonymity greatly contributes to the virality of manipulated content. There is (in many cases) a wrong assumption that anonymous internet behavior is free of consequences. On top of that, anonymity serves as the ground for increasing dominance of trolls in cyberspace [89]. An additional lately increasing behavior is protesters worldwide who wear masks as means of “real life” or “offline” anonymity. This is mostly done to avoid the large screening of faces through large journalist coverage as well as avoiding smart cameras that are integrated in many cities and allow face recognition<sup>3</sup>.

### Decentralization

Decentralization is a technological concept that influences different means of propaganda and terrorism, e.g. the way to convey messages or financial organization. Decentralized financial activities with cryptocurrencies are out of the scope of this thesis, but they play a vital role in funding terrorism by benefiting from blockchains and global resources of money [127].

Technological decentralization refers to the distribution of data, multiple operations of data processing and storage by different independent entities, and the lack of a single point of control that governs the activities or decisions. Decentralized technologies, such as peer-to-peer networks, enhance the resilience and security of systems, while preserving privacy and reducing the ability of a central authority to control data flows and thus censorship. The increased autonomy of the entities in decentralized chains can also be largely misused, with uncontrolled criminal activities (as often seen on the dark web with the Tor network), distribution of offensive content or mis- and disinformation that is hard and in times infeasible to regulate, and a broad attack vector, including sybil attacks in which multiple fake identities are overloaded to the network by a single entity getting more control over the network, consensus manipulation, routing attacks in which the propagation of messages is manipulated, distributed denial of service (DDoS) with the flooding of requests by a single entity, etc. Thus, the technological structure of decentralized systems has direct societal implications [3][140].

We have seen different examples of centralized propaganda in the past, namely pro-

<sup>3</sup><https://www.nytimes.com/2024/05/02/nyregion/college-campus-protests-anonymity.html>, Accessed on: 01.06.2024

paganda that stems from one source, mostly some totalitarian regime that controls all means of communication and “traditionally” distributes the content. This type of propaganda is mostly done in an isolated environment, where individuals are only exposed to one narrative that the people in power dictate and extremely limit the freedom of speech. Robinson and Whittaker [153] mention the 12-week-long Rwanda Genocide in 1994 in which about one million people (mainly Tutsi) perished. It is a clear example of one-channeled propaganda, done by radio. It is strongly believed that the wide and brutal violence exercised toward this group of minority, which was extremely “othered” based on different historical events, was an outcome of the fuelling hateful multiple broadcasts over the radio channels [12].

Another example of centralized propaganda is the Nazi propaganda against Jews (again using the “othering” approach and parallel creating and strengthening the community feelings of the German society, the “Volksgemeinschaft”) before and during the WW2 that used various media channels to promote their ideology and the so-called “Weltanschauung”. Joseph Goebbels, the minister of propaganda and public enlightenment, mainly dictated the tone of the propaganda. The utilized media included newspapers, films, radio, and posters. The intensive propaganda was accompanied by aggressive exclusion of opposing opinions<sup>4</sup>

From the above, we learn that the nature of propaganda transformed into a decentralized way of distribution and it stems from the enabling large technological infrastructure with a wide-range of features and platforms for communication. Multiple sources (some of them might intersect) promote various ideologies as part of the decentralized propaganda. This way, it reaches a large and diverse audience overcoming national borders. It mainly gives the fake feeling of having a multi-faceted picture and the ability to “critically think” is developed enough under these conditions to decide what is the truth and what to reject. In reality, this whole phenomenon only contributes to the post-truth era, rejects the traditional media as a reliable source of information, and the feeling of believing in nothing and in everything at the same time is intensified [95]. Therefore, there is no one line of thought or Weltanschauung that we can follow or reject, as happened during totalitarian regimes and in the examples above.

### 4.3.3 Algorithmic Extremism Detection

The rise of online extremism has urged new research directions aimed at detecting extremist activity or misuse of technologies, particularly given the vast amounts of information that only automated processes can efficiently manage. These technologies are adopted by law enforcement in their anti-terrorism efforts, but also by social media platforms seeking to improve the safety of their digital spaces by strengthening their content moderation. Detection techniques of extremism and violence become vital safeguards for generative

---

<sup>4</sup>United States Holocaust Memorial Museum. “Joseph Goebbels.” Holocaust Encyclopedia. <https://encyclopedia.ushmm.org/content/en/article/joseph-goebbels-1>. Accessed on 16.05.2024

AI systems, which are susceptible to extremist misuse.

Extremism detection algorithms mostly focus on textual content and visual content. These include predefined keywords to flag harmful content or more sophisticated machine learning algorithms, primarily sentiment analysis for textual content, or feature extraction for visual content. Govers et al. [67] emphasize the absence in research of cross-examine techniques and models for the detection of extremism, hate speech, and radicalization. The relevant solutions to the problem of generating violent visuals, involve both techniques for extremist input detection and visual violence detection for the potential harmful outputs.

### Textual Detection of Online Extremism and Hate Speech

Detecting hate speech on social media has obtained the primary focus in extremism detection algorithms. Textual analysis stands at the core of the fight against online radicalization. Defining hate speech is non-trivial, as humans share different levels of intolerance toward extremist views and tend to observe text in a subjective manner. Analogously, the digital representation and labelling of hate speech is a difficult task, that aims to reflect societal values while understanding language nuances [67]. We focus on two widely used algorithms present in our working corpus [67] [61] [7], which are often employed to detect hate speech and extremist textual content, namely keyword detection and sentiment analysis. These involve large language models (LLMs) that include multiple languages to target different audiences.

**Keyword Detection.** Keyword detection is a foundational technique in Natural Language Processing (NLP), widely applied in search engines, content moderation, and social media monitoring. It serves as an important base for more advanced NLP methods. The keyword detection process typically involves specific words or phrases within an examined text or corpus, often using a combination of language models and keyword extraction methods [90].

The simplest keyword extraction methods include naive rule-based approaches, often relying on predefined keyword dictionaries. Named Entity Recognition (NER) is a more specific application of keyword dictionaries, that is especially useful for tracing extremist content related to known figures and organizations. These dictionaries are frequently constructed with human oversight, especially in domain-specific tasks where expert knowledge is needed [90].

More advanced keyword extraction methods analyze the importance of the looked-up term against a given corpus by applying statistical methods and are therefore more optimal for automation. The two most used methods are N-grams and Term Frequency-Inverse Document Frequency (TF-IDF) [7]. N-grams capture the immediate context of a single word, by examining  $N$  sequence of words as a single atomic unit. Bi-grams and tri-grams (two or three consecutive words, e.g. “social

media”, or “streamed terror attacks”) are frequently observed, due to exponential growth of possible word combinations. TF-IDF calculates the importance of a looked-up word by comparing its frequency in a given document (the frequency of the word in the document) against its frequency across the whole collection of documents that build the corpus of the given language model. This calculation gives words that appear frequently in one document but rarely elsewhere a higher weight, flagging them as significant. This enhances the keyword detection’s ability to focus on contextually important terms. These methods enhance keyword detection by observing its meaning within a context, which is crucial in understanding linguistic and language-specific nuances [90].

In the context of hate speech detection and extremist content, keyword extraction is often based on indicators drawn from literature, integrating radicalization studies and social science studies. Further sources for harvesting relevant extremist keywords are social media platforms, especially pages and chats known to be used by radicalized groups and individuals. These constructed databases leverage current expert knowledge. However, researchers have found that metrics based solely on keyword extraction, also if well-performing, present limitations as standalone methods [7] [61].

**Sentiment Analysis.** Sentiment analysis focuses on examining the emotional tone behind textual content, expanding the scope beyond keyword detection by identifying the writer’s sentiment, with the classic categories of positive, negative, or neutral. Often, sentiment analysis is applied in a domain-specific manner to better capture context-based nuances. The two most common approaches are lexicon-based and machine learning-based methods, with hybrid approaches combining both strategies [90].

Lexicon-based sentiment analysis involves predefined dictionaries, similar to keyword detection. These lexicons are annotated with their corresponding associated sentiment scores. The algorithm scans a given text, identifying the sentiment of words that match the lexicon entries. Sentiments can range from simple labels to deeper emotional annotations, e.g. sad, fear, or anger, and may also include the different intensity levels of a given sentiment. The success rate of this method heavily depends on the quality of the lexicon and its compatibility with the domain-specific language [61] [90].

More advanced sentiment analysis methods use machine learning algorithms to optimize sentiment detection. A common approach is the probabilistic Bayesian classifier, which makes a naive assumption that the features (words) are independent of each other. It applies the conditionally independent probability calculation of the Bayes’ theorem as follows:

$$P(\textit{sentiment}|\textit{words}) = \frac{P(\textit{sentiment}) \times P(\textit{words}|\textit{sentiment})}{P(\textit{words})}$$

This calculation is simple, efficient, and interpretable, thus enhancing explainability. However, the naive assumption about the independence of the words often misses the actual contextual meaning, struggling with linguistic nuances like negations or sarcasm, or sentences containing multiple sentiments. Furthermore, the Bayesian approach is susceptible to bias, especially when the training data is skewed toward entities with higher representation [61] [90].

Deep learning models outperform simpler approaches in sentiment analysis tasks. Two primary models are Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT).

LSTMs are complex neural networks that capture context by processing sequential inputs and producing contextualized sequence of outputs. They are designed to filter irrelevant information and focus on relevant details. Gates within the LSTM architecture control information flow between layers, ensuring context-aware decision-making, and incorporating mathematical functions between the layers based on the gate's type. BERT is an unsupervised deep learning model based on multi-layer transformers.

BERT's architecture involves two steps; pre-training and fine-tuning. In the pre-training step, unlabeled data is trained over different tasks with masked sentences with the aim of filling in the blank space. In the fine-tuning step, the pre-trained model integrates the repeated task-specific feedback in a self-supervised manner (no human-annotated labels are involved). The fill-in-the-blank task is according to the Masked Language Modeling (MLM).

Convolutional Neural Networks (CNNs) are commonly used in visual tasks, but they can also be adapted for sentiment analysis. CNNs consist of three main types of layers; convolutional, pooling, and fully connected layers. Each layer performs specific mathematical operations to process and propagate information through the network. In sentiment analysis, CNNs treat input text similarly to how they process images. Instead of pixels, CNNs treat words or word embeddings as input features, capturing local patterns in sequences (such as phrases or word combinations) that can help determine the sentiment of the text. This makes CNNs effective at identifying important features in text, much like they do in images. [42] [61] [90].

For hate speech tasks, the pre-training phase often incorporates a corpus of frequent extremist words classified by researchers as extremist or non-extremist content. Word embeddings are often used to cover a larger space of semantics. Frequently used classifications include labels of neutral, moderate, low extreme, and high extreme, but also a wide range of emotions that are all negative such as anger, fear, and sadness. Most of the research approaches divide extremist corpora into

different ideologies, optimizing results through specific terminology [61].

Keyword detection and sentiment analysis serve as key components in the detection of textual content of online extremism and hate speech, offering complementary approaches to understanding both the explicit and implicit emotional tones in textual content. While keyword detection efficiently identifies known extremist terms, sentiment analysis captures the underlying emotional context, making it essential for flagging more nuanced and potentially harmful communications. These methods form an automated combating mechanism against online radicalization. The continual integration with domain-specific knowledge is necessary for optimal accuracy and relevance.

### Visual Recognition of Violence

We focus on violence detection in images and videos, which is particularly relevant given our concerns about generating extremely violent content with generative AI systems. Besides social media platforms that apply removal policies on violent content, this topic is also critical for smart surveillance systems, which aim to develop automatic methods for action and activity recognition. Detecting visual violence is fundamentally a classification task concerning the object and the activity, typically employing deep learning techniques like LSTMs and CNNs, which are also used in sentiment analysis, as we covered in 4.3.3. These models rely on datasets of violent visuals, including material from movies, YouTube videos showing violent crowds, or surveillance footage capturing real-life violent acts. For example, Sultani et al. [174] published the UCF-Crime dataset, a publicly annotated resource supporting only recognition tasks of real-world crimes [187] [146].

However, Mumtaz et al. [125] highlight several technical challenges in using deep learning algorithms for violence detection in videos. The complexity increases when analyzing sequential frame data involving time-series patterns, as both spatial and temporal features must be considered. Violence is often detected by recognizing anomalous patterns of activity, such as unusual or erratic movements. In this context, LSTMs are particularly effective at analyzing sequences of frames, while CNNs are useful for extracting features to shed more light on the domain of the videos without prior knowledge that is often necessary. Razman et al. [146] promoted the hybrid approach that combines both the LSTM and CNN architecture. Mumtaz et al. also emphasized the limitations in the current capabilities of deep learning algorithms, particularly in the range of attributes they can classify.

### Technical Challenges of Detection Algorithms

A key challenge in developing violence detection algorithms is the subjective nature of defining what constitutes violence, having a subjective nature. This could lead to inconsistencies in datasets or results of validation processes [146] [125] [67]. Moreover, violence recognition extremely depends on available data which may introduce significant biases if training datasets are skewed or not diverse [125]. A further challenge is the

intentional bypassing of malicious actors, who study the detection mechanisms and develop techniques to circumvent them, often with the help of IT experts. We will further discuss the structured terrorist misuse of technology in 4.4.2. Additionally, privacy concerns arise when harvesting large-scale training datasets, as these systems could potentially be misused for mass surveillance, undermining the safeguards originally intended.

The technology component with all of its elements greatly eases access to extremist content. Furthermore, the feeling of moral outrage gets intensified by the consumption of “first-hand” visions of brutalities. This reached emotional state is also an integral component in the radicalization process that eventually leads individuals to be self-radicalized and construct extremely polarized opinions thought to be based on the fully explored picture that reflects reality. This assumption has also been largely studied; as explained by Sullivan [173], the internet serves as a breeding ground for shaping and reinforcing violent extremist narratives that finally enhance and amplify the radicalization process of the individuals. Detection algorithms designed to identify extremism, hate speech, and visual violence can serve as critical safeguards online. Moreover, they can be integrated into generative AI systems to prevent the creation of violent content, both by flagging problematic prompts and inhibiting the generation of violent visuals.

## 4.4 STRUCTURE: Social Media and Terrorism

The second social component in our concerned socio-technical system is the organized and structured misuse of the internet infrastructure and social media platforms by terror organizations and extremist groups. In our corpus, we found 19 papers that contribute to understanding this component. The algorithmic and socio-technical structure of social media platforms that is often AI-powered enhance the social phenomenon of radicalization processes. The structure component exploits these two first components and builds its layer of organization on top of them. To understand the structure, we firstly need to understand the definition of terrorism, how its organized activity has transformed with the development of the internet infrastructure, and how social media platforms have facilitated a whole new playground for merging online and offline terrorism. The subject of counter-terrorism is out of our scope and is therefore not discussed.

### 4.4.1 The Roots and Danger of Terrorism

The nature of terrorism poses a novel threat to individuals and democratic societies as well as infrastructure [46] while continuously learning new ways to cause harm [50]. In his paper, Primoratz [141] tracks down the definition of terrorism by which repulse the majority. “Terrorism” derives etymologically from “terror” which means *extreme fear* and the action of terrorism utilizes such extreme fear as a means of creating intimidation with the additional layer of an accompanying purpose. Some ideology wraps the purpose and it serves as the driving motive. Ideologies are in nature binary colored, namely, rejecting a fine-grained approach to problem understanding and problem-solving. This total approach

gives many the certainty they seek, by selling “clear” and simple narratives, as Webber [185] explained. This also helps in conveying ideologies to the masses, as the individuals do not have to critically engage with the topic, as it is provided in such a simplistic manner.

Laqueur [99] points out in his book that throughout the history of terrorism, the nature of it has changed from one era to another as well as from one area to another but one thing stayed common, namely, it has mostly failed in achieving the cause that is fought for of changing a political situation. Furthermore, the structure of terrorism is based on two targets;

- *The indirect target and of primary importance:* the cause and the future it aims at achieving and that includes this group of people that are being forced into behaving in a certain way that is achieved through terrorist pressure
- *The direct target and of secondary importance:* those who are affected by the direct violence exercised to reach this target and consist of innocent people.

The additional individuals who are prone to engage in exercising terrorism is also a key component and which we discussed in 4.2.2. Engaging in terrorism, planning, and executing terrorist attacks is to 95% of the cases done in groups demonstrating the significant role of group belonging and group identity [45] that also was shown in 4.2.3 to be a crucial component in the radicalization processes.

Terrorism represents the most extreme manifestation of violent extremism. Risius et al. [150] presented a four-level ideological engagement that includes a continuum ranging over individuals who fall under the categories of partisan, fringe, violent extremist, and terrorist. Those who belong to the violent extremists group exhibit cognitively intolerant toward other groups, exclude them, and view their moral obligations as inferior. Behaviorally, they systematically engage in dehumanizing others and employ means that harm the physical and mental well-being of others. Terrorists, a subset of violent extremists, escalate this pattern by cognitively color target groups as the “enemy”, even legitimizing their death. They additionally intimidate the general population through the conveyed “message” and behaviorally endorse and commit acts of physical violence, often accompanied by glorification of willingly martyrdom.

This work considers both violent extremism as well as terrorism in modern times while the main focus is on structured extremism that terror organizations and extremist groups exercise and promote. We acknowledge the misuse potential by populists or nationalists but we prefer to limit our scope based on higher-risk actors and lower-risk actors, based on mortality in modern times. Terrorism encompasses not only physical violence but also cyber and psychological violence.

#### 4.4.2 Online Terrorist Infrastructure

The online infrastructure serves terror organizations and other extremist groups as a tool for recruitment, dissemination of propaganda, intimidation, facilitation of terror attacks, and transformation of their communication strategy and mobilization, regardless of the geographical boundaries [173]. Terror organizations and extremist groups have long recognized the potential of leveraging internet infrastructure and social media to promote, expand, and coordinate their networks [11]. There is also a strong correlation between the time spent by individuals on social media platforms and also their level of engagement online and the success of these aims [80] [33].

Content provided by terror organizations poses a great risk to society by abusing the freedom of speech. It reaches not only the masses but the personalization of the internet and social media platforms targets relevant individuals that belong to the same filter bubbles which accelerates even more the networking of radical groups online. San Biagio et al. [161] refers to multiple studies that confirm the prevalence of organized terrorism on social media platforms. These studies highlight the use of the internet by terrorist groups to disseminate their propaganda and recruit new members. The online propaganda, which is part of the last socio-technical component, targets often the offline world. Another usable social media feature is the use of hashtags on shared content which terror organizations have also largely exploited to target individuals [161] [171].

The first terror organization to exploit the internet for their communication strategy was Al-Qaeda which launched the 9/11 attacks and planned it via encrypted messages between Afghanistan and terrorist cells located in the US, as UNESCO's report from 2017 mentions<sup>5</sup>. We can further learn from this report about the Islamic State's (ISIS) use of the internet. IS separated from Al-Qaeda in February 2014, and became the first terror organization to control a territory with the full scope of institutions as found in legitimate governments that has managed to lure people from all over the world [139]. Social media has played an intrinsic role in ISIS' rise and success [117]. Part of ISIS' methods to achieve worldwide acknowledgment is due to its globalized approach, namely, the spread of material is multi-lingual reaching potential individuals also outside of the Arabic-speaking world, as Pashentsev et al. [137] discussed. Far-right extremist groups have also been widely present throughout the internet with their own websites and forums. Part of their techniques to disseminate propaganda is including in their website attractive components of music or video games accompanied by disinformation and their unique language as part of their communication strategy [75].

Social Engineering techniques are also common in online radicalization and extremist groups and terror organizations often employ them to target individuals. These techniques include *authority* (when gaining information), *conformity* (in accepting a

<sup>5</sup>[https://ec.europa.eu/programmes/erasmus-plus/project-result-content/b63fe787-88d2-443e-a9ac-e0adbac2e214/UNSC-Study-Guide\\_final.pdf](https://ec.europa.eu/programmes/erasmus-plus/project-result-content/b63fe787-88d2-443e-a9ac-e0adbac2e214/UNSC-Study-Guide_final.pdf), Accessed on: 18.05.2024

group behavior and normalizing radicalized thoughts or behavior), *liking* (the feelings of being liked that fringe individuals receive from a community), *commitment and consistency* (pressuring individuals into a situation and assuring it is maintained) *reciprocation* (the bidirectional exchange of acts), and *distraction* (creating an increased emotional response to the point of no logical decision making). Applying these methods take advantage of the existing societal tendencies and radicalization processes in an organized and targeted manner [158]. Organized radicalization efforts are often successful when an extremist group manages to fulfill the needs of the targeted individuals and for this cause they often disseminate manipulated information to deceive [122]. They often address the general need to belong, which a virtual community can satisfy, or have a meaning from a presented ideology, which are fundamental elements of radicalization models. They manage to address further unique needs, based on the digital profile of individuals [117].

To facilitate social engineering techniques and further methods online, ISIS, as a specified example, had approached IT experts that helped the organization's lifecycle in different sectors. The expertise is evident, for example, in its ability to rapidly create new accounts after social media platforms remove some [139]. As part of their strategy of disseminating propaganda online, they have developed their own social media platform, passing by the security mechanism of the common platforms and having no restraint or monitoring on their content. It is a clear use of the decentralization concept discussed earlier in 4.3.2.

Recruitment is one of the main activities of terror organizations and other extremist groups over social media platforms. There have been multiple evidence for online organizing are major for recruitment [188]. For instance, there is an estimation of about 30,000 foreign members who physically joined to ISIS in Syria after making an online contact who later participated in ISIS' deadliest attacks. These individuals are also considered to be more dedicated to the ideology and the cause and thus engaging more in risky attacks than local members. One frequent recruitment method is "love bombing", namely, direct contact to individuals who engage with radicalized content of the group [43]. Saad and Alhumaid [157] described this method as a "knock on their doors" and emphasized the difference to the traditional internet sites without the features of prompt engagement with the content in which individuals had to land on the specific group's website. Video posts are also highly effective in convincing individuals to join with striking statistics of a fourth of foreign recruits being mainly influenced by content from YouTube [43].

Mobilizing individuals is another common tactic of online terrorism that aims to coordinate offline activities in an organized manner, e.g. encouraging participation in demonstrations or rallies as well as acting violently, and its identification by the authorities gets obfuscated by the high volume of online communication and content [35] [172] [26]. Esmailzadeh [50] addresses another mobilization-related risk, namely the ability of AI algorithms and available big data to analyze the vulnerability surface of future attacks and plan attacks accordingly to reach a maximized damage.

Terror organizations and extremist groups have expanded their activity from offline spaces to the digital realm. They manage to exploit the inherent radicalization mechanisms of social media platforms by targeting potential individuals based on their preferences and online activity. Their primary organization-related activities encompass recruitment, the dissemination of propaganda, mobilization, all of which are significantly facilitated by the available online infrastructure in a sophisticated manner that largely includes deception and misinformation.

### 4.5 TASKS: Propaganda, Polarization, and Psychological Harm

The second technical component, the task, and the final element of our socio-technical system stands for extremist and terrorist activities and their malicious goals. These activities build upon the tendencies in society to engage with radicalization of thought and behavior, the role of social media platforms in connecting individuals from the mass crowd with similar ideologies and contributing to self-radicalization, and the efforts of organized extremist groups and terror organizations to exploit the internet infrastructure to lure prone individuals and recruit them.

In this component, we gather together the different activities and their consequences that involve extreme visual violence and primarily relevant to radicalization, polarization, and psychological harm which are highly relevant to our research question (RQ1). The focus point graphic violence is due to technological capabilities of generative AI models to output manipulated visuals that can amplify scenes of visual violence. The task in the context of extreme visual violence is a compound of common visual propaganda that integrates extreme violence, also named “propaganda of the deed”, political polarization as a secondary and cognitive consequence of a successful radicalization process, and psychological harm of a secondary trauma that the exposure to extreme visual violence causes following an extremist activity, also known as psychological warfare. The propaganda is part of the middle component of the radicalization pathway, namely the radicalization process itself, and polarization and psychological harm belong to the last layer of the pathway. The understanding of these activities is a fundamental component in portraying the risk landscape of terrorist and extremist activity with the application of generated and targeted visual violence.

#### 4.5.1 Extreme Violent Visual Content as means for Propaganda

*“A picture is worth a thousand words”. And what about a video?*

Images and the visual component of videos are language independent and tell “one clear truth”, hence can reach and convince large audiences. Additionally, social media has become an integral component of modern daily life that reshapes reality, as Küey [96] claims, as it does not necessarily reflect reality but rather reconstructs it. Consequently,

acts of terrorism and the nature of armed conflicts have long left the sheer physical dimension whilst the virtual world is no longer a distinct realm as it used to be. The large extent of journalist coverage in the earlier years and later the emergence of social media and the shift of extremist content to those platforms greatly helped this tendency to take rise.

Social media has brought along the “Everyone is a Journalist” phenomenon, namely, the role of the media in the traditional mediation of reality to the masses has been taken by everyone with a camera and internet connection and the feeling of “Liveness” and authenticity intensifies the experience of the viewer. This also means that the traditional intervening or monitoring done by the institutions no longer exists when it comes to individual reporting and this role was passed on to the social media platforms that host and enable the distribution of the content to the masses.

Simultaneously, the traditional media focus more and more on sensational news to lure readers and watchers and put at the front information on violent events [119] [16] [8] [148]. Another outcome is the ongoing “post-truth” era, due to the flow of news that individuals consume, is examined by means of acceptability or believability and no more by accountability or reliability [96]. Cambridge’s definition of “post-truth” includes the state in which individuals are more likely to accept arguments based on emotions and beliefs and not undoubtedly on facts<sup>6</sup>. The fact that there is an increase in the spread of misinformation on social media platforms does not contribute to the credibility and reputation of traditional media. On the contrary, there is a rising tendency to strongly doubt any information coming from these communication channels [32]. Social media platforms have transformed into a battleground to convey narratives and their authenticity. Parallel, extremist content has become more present in mainstream virtual spaces [95].

The horrible beheading videos of ISIS (among other extreme violent acts) starting in the year 2014 have marked the beginning of a whole new era of visibility and “aesthetics” of violence targeting the large global crowd [8]. Rose [154] speaks about ISIS’ unique visual aesthetic that serves as a signature of the terror organization, together with their distinct colors and the use of the logo in all the fetishized HD visual materials, creating an immediate “brand” recognition effect [181].

Similarly to ISIS’ visual aesthetic, Schulte-Sasse [163] speaks about the Nazis’ distinct visual style that correlated with the fascist ideology and was seen in diverse fields, e.g. cinema, posters, architecture, and uniforms. Thompson [137] explains that this aesthetic drives the communication strategy by relating extreme violence to the visuals of action films or computer games. The attempted censorship in this case did not help, and these videos are still widely spread on different social media platforms and especially the dark web. De Zayas and Matusitz [41] found a clear contribution of ISIS brutal

---

<sup>6</sup><https://dictionary.cambridge.org/de/worterbuch/englisch/post-truth>, Accessed on: 16.05.2024

videos to the establishment of multiple communication channels in a decentralized manner.

Later, as Ibrahim et al. [83] addresses in their work, the horrific case in Christchurch, New Zealand, in which a live-streamed massacre of 51 Muslims in two mosques by a far-right white supremacy terrorist took place in 2019, marked a new record low while utilizing Facebook’s infrastructure. The New Zealand government promoted a law banning the streaming of the cruelty with a punishment of up to 14 years in jail [109].

The phenomenon of bloody images and videos being widely spread and used as propaganda itself emerged in the last couple of years and is considered to be one of the most effective propaganda types [134]. It is also called “Propaganda of the deed”, as introduced by Kaplan [91] and Thompson [177], or “Atrocity Propaganda”, as referred to in Baugut and Neumann’s work [16]. It is a common practice as well as a main component of the communication strategy in Islamic extremism. The visibility of this type of propaganda is especially eye-catching, particularly due to the bloody scenarios that might seem appealing to certain individuals that convey the message[10]. This emphasizes even more the bidirectional influence of technological developments and social phenomena.

A very clear contrast to the of the “Propaganda of the deed” can be seen back in the times of WW2 in which the Nazis did everything they could to hide any evidence about their brutal deeds from external audiences. This was seen even in the ways of extermination, using crematoriums and putting on fire many of the ghettos and concentration camps to destroy any hints about the mass killing<sup>7</sup>. Nevertheless, during the attack on Poland in 1939 by Hitler’s dictatorship, the terrorist methods included visuals that were related to violence that the Germans exercised as a war tactic [177].

More recent cases of brutal violence being largely distributed and discussed on social media platforms include the terror attack in Vienna in November 2020, the Ukraine-Russia war, and the conflict zone in Israel and Gaza starting with the October 7th attack. The Vienna case is particularly interesting, as it was the first terror attack to had happened within 39 years in the city and there was a high participation from the public in stopping sharing the video of the shooter in action hence cutting the viral chain. Furthermore, the press institutions obtained more than 1500 official complaints about publishing the videos and images from the attacks [106].

In contrast, that was not the case for the October 7th attack launched by the terror organization Hamas in Israel from which a high amount of brutality against civilians became viral, also live-streamed content, including graphic content of abusing people and corps, was shared online on various platforms, social- but also traditional media in different grades of censorship by all sides and worldwide. This attack brought with it a whole new availability to extreme visual violence shared not only on side channels

---

<sup>7</sup>United States Holocaust Memorial Museum. “Nazi Propaganda.” Holocaust Encyclopedia. <https://encyclopedia.ushmm.org/content/en/article/nazi-propaganda>. Accessed on 6.4.2024

on Telegram but also on mainstream social platforms [102]. This large exposure to such visual violence raised major concerns regarding the national mental health among specialists in the country leading psychiatrists to widely warn the public from watching them as well as the Ministry of National Security for Child Protection in the cyberspace of Israel publishing a national warning statement in Hebrew and Arabic<sup>8</sup>.

Extremist groups and terror organizations often employ bots to disseminate propaganda and misinformation rapidly. Al-Khateeb and Agarwal [4] examined how bots were used to disseminate ISIS' beheading videos, finding different bots behaviors, including their posting and re-posting activities. They found that most of the bots had similar account names and used in their posts unusual characters. Moreover, most of the originated propaganda from platforms other than those where it was widely disseminated. By exploiting the virality mechanisms of major social media platforms, these bots managed to amplify extremist content.

This whole new pool of atrocities is a potential source of training data for Generative AI systems and by understanding the full scope of the available content, we can modulate the potential risks that come along with such violent output and understand how the nature of modern terrorism changes together with the technological developments. We acknowledge the risk of having both real and fictional extremist violent visuals and follow Sageman [159] claims that the exposure itself to extremist content can incite the viewers the feelings of moral outrage that leads them to take action among other accompanying psychological and social aspects.

### 4.5.2 From Pluralism to Polarization

The interplay between pluralism and polarization has never been more relevant, in an era of immediate access and large distribution of everyone's word and world. The term "Pluralism" is in essence a way to have social diversity, namely, different groups in society respect their differences and demonstrate tolerance toward them. According to Yumatle [189], it stems from political philosophy and has been a vital component of democracy already in Ancient Greece. The different kinds of pluralism consist of cultural, political, and philosophical and reflect the coexistence of different ethical values, worldviews, and practices. The known sentence "I disapprove of what you say, but I'll defend to the death your right to say it" that was written by Hall about Voltaire's beliefs [36] demonstrates the concept of freedom of speech within a pluralistic society, the equal inclusion of fairly different opinions in society, and the importance of having diversity. Baldassarri et al. [13] speak of a model of political pluralism that is the base for democratic systems that consist of mutual interests and identities while enabling access to political representation for most groups.

Following, we will see how polarization takes this diversity and tosses the inclusion

---

<sup>8</sup>[https://www.gov.il/\(he|ar\)/departments/news/warning\\_105](https://www.gov.il/(he|ar)/departments/news/warning_105), Accessed on: 17.05.2024

component away while emphasizing the divergence to the point of no return. Fiorina [55] refers to polarization in the political context as opposing ideologies (including principles, tendencies, and world views) that create a “bimodal distribution” of views and political preferences. Further explanations refer to polarization not only as radicalized opinions, but also the process of growing ideological divergence within a society, expressed through public opinion, and the enhancement of division between political groups, as Baldassarri et al. [13] discussed.

Political polarization contributes to a democratic breakdown and even to democratic backsliding by posing a real threat to the stability of democratic societies [34]. This is done by shifting ideologies from the center toward the extreme to the point that it is no longer possible to bridge the gaps of the non-overlapping views. Dalton [40] also named this type of polarization *ideology polarization*. The second type of polarization is *affective polarization* which puts in focus the role of identity in a political context and the level of antagonism and even dehumanization toward out-groups [113]. McKeown et al. [118] showed in their analysis that dehumanization and distrust dominated the public journalist discussions following a case of a terror attack which emphasized the political polarization toward the out-group. Callender and Carbajal [29] explain that the moderate opinions that slowly vanish and the created gap between extreme views, also called the “center”, play in favor of extreme parties in their political competition. Abramowitz<sup>9</sup> named this process the “disappearing middle”. At this point in which the gap is too big, the voters cannot identify with the party from the other side of the political spectrum and it fuels those parties to become more extreme and polarize more to lure more voters without fearing to lose them. This phenomenon happens iteratively as a feedback loop and the gap between the different spectrum edges gets deeper and deeper incrementally.

On the one hand, polarization plays a role in times of war, extremism, and terrorism, and on the other hand, such times can also bring people together and create social resilience. These phenomena help understand what happens psychologically and socially after being exposed to extreme violent scenes.

### **Polarization in Times of War, Extremism, and Terrorism**

Conflicts and times of war have always had many facades and contrary world views clashing with each other. The sociologist Durkheim [48] referred in his approach to disruptions in social solidarity and common consciousness as periods of confusion within society which he named as anomie. During those, individuals are more prone to radicalization and polarization while having, according to Durkheim, too much freedom and no clear frame about judging what is right and wrong. The work of Risius et al. [150] that we referred to earlier in 4.4.1 modulates a dynamic matrix of extremisms and terrorism. The two first groups in the 4-stages matrix that we have not touched yet are the *partisanship* and *fringe groups*.

<sup>9</sup>Han, Hahrie. “The disappearing center: Engaged citizens, polarization, and American democracy.” (2010)

The partisanship is especially of interest to the topic of polarization, as it is the first layer of the extremism scheme that eventually leads to a polarized society. Partisanship states the strong sense of association with a political party that brings together individuals with the same norms and ideologies. Partisanship is different than mainstream world views, as it creates a social identity within the support network whilst opposite views are dismissed. Additionally, the group dynamics include a common political grievance and are sometimes accompanied by the search for significance.

Fringe groups take partisanship to the next level, namely, they are still supportive of non-violent ideologies but are on the edge of society and big social structures. Further elements that go along with the fringe groups are the glorification of the in-group members and the discrediting of the out-group. In some cases, there is an isolation from the general society. Furthermore, conspiracy theories and factors of internal and external parties to blame are promoted. Another element of this group dynamics is the censorship of confronting views and the in-group indoctrination into binary thinking [150].

The base tendencies toward polarization put focus on the in-group vs. out-group identity (also known as “othering” and “us vs. them” rhetoric), tunnel view that disregards opposing ideas and the created simplistic way of observing and interpreting the world that allows extreme ideologies to be largely conveyed. Polarization does not merely stay in the political dimension and related social issues handled by politics and it has a deep social effect. Makrehchi [112] researched the correlation between language shift and social conflicts in the current polarized social media. In his work, he speaks about the day-to-day interactions that are characterized by strong and extreme terms evoking correspondingly extreme emotions and sentiments. In polarized societies, there is a high tendency to engage in controversial topics that receive the main spotlight and create more division. The vocabulary involved in such a polarized environment is distinct and consists of labeling and tagging of the out-group. The change of used language eventually leads to cases of racism in society, especially because polarization vocabulary emphasizes the superiority of the in-group vs. the inferiority of the out-group. Such cases include violence, hate crimes, and (state and non-state) discrimination.

The hostile environment within a polarized society following these cases decrease the level of civil courage and the feelings of mutual trust and responsibility that are necessary for the continuity of democratic societies, as McCoy et al. [116] comprehensively overview. Rychwalska et al. [156] demonstrate with their model that social media gives the platform to enhance polarization through the divisive dynamics of individuals combined with the described filter bubbles.

We have already seen in 4.5.1 that visual content provides an immersive experience for the viewer which results in intensified feelings. Those feelings can be easily manipulated by different parties to serve their needs. We have seen that strengthened social

identity and the blame of out-group members for the reasons of grievances or injustice is a fundamental component of polarization. We can conclude that having visual content involved gives a whole new dimension of trustability and reliability to the different arguments. The viewer obtains the feeling of control when it comes to judging situations that are supposedly based on “visual facts”. This way, violent visual content toward the in-group can be exploited, hence the feeling of moral outrage gets amplified while having such “strong evidence” about the threat of the out-group to the in-group.

### **Social Resilience after Exposure to Wars and Terrorism**

In times of social disruption, especially by external factors, such as wars and terrorism, one can also observe the opposite phenomenon of polarization, namely, bringing the hearts closer together, which we will look at next. Collective traumata, and especially terrorist attacks, create a safe space and environment for the different members of the group to come closer together and express their emotions, according to Durheim’s theory [47]. A highly linked phenomenon is the “rally around the flag” which describe the short-term surge in popularity and widespread support for government representatives during times of war or crisis [64]. It leads to a higher degree of solidarity among these communities and refreshes in a way their feeling of belonging and the shared beliefs that are dominant in the group.

This has an overall positive effect of community support, strength, inclusion, and tolerance of other individuals, accompanied by emotional synchronization of the group members [110]. Garcia and Rimé [62] tracked this phenomenon and referred in their work to the 9/11 attacks, the Madrid attacks in 2004, and the Charlie Hebdo attack in Paris in 2015. They discuss the immediate increasing need to talk among survivors, also observed as the “social sharing of emotion”, and that it also seems to decrease in the long-term, as it happened in the case of the Madrid attacks in 2004 after two months. These collective emotional episodes, as Garcia and Rimé named them, result in emotion-sharing feedback loops through exposure to news and the accompanied social interactions.

The higher degree of solidarity and connectedness feelings, according to Cacioppo, Reis, and Zatura [28], are vital to social resilience, which is the general capacity of the population to undergo stressful situations and recover from them while sustaining positive social bonds between the group members. Social resilience puts front the capacity of individuals to work with each other towards finding in tragedies the opportunity to grow out of the trauma and turning the hardships into an advantage. It also emphasizes the importance of interconnections in the population and its contribution to the collective development, learning, and growth, while preserving non-monolithic views among the group members and benefiting from the differences. Social resilience is based on pluralism in the sense of emphasizing the common factors and dismissing the negativity that characterizes polarization.

Community resilience takes place in different settings. Brajawidagda et al. [24] dis-

cussed urban resilience that describes the ability of the people within a city to recover as a community from a hazard. Their work highlights social media as a major contributor to various components of urban resilience, while emphasizing the importance of competence (actions, skills to solve problems and community empowerment) as a vital factor in fostering resilience. Hamiel et al. [71] considered municipal programs to strength urban resilience in times war and hazards as part of a disaster risk management strategy. Their aim is to improve the population's functional and social skills to deal with stressful events and trauma.

Exposure to visual violence from terror attacks or war zones brings closer these horrible events to the mass crowd, especially in times of social media in which such content is largely distributed and has turned into an integral part of the daily use of such platforms. The exposure itself can trigger a secondary trauma, but can also create an individual resilience to visual violence (also called media violence). Individual resilience, and in our case to violent visuals that are more and more common, is crucial for well-being and mental survival and refers to the ability to continue functioning after going through adversity [73]. The individual's resilience is necessary for a functioning society, as it eventually effects the community resilience and on a larger scale the societal resilience. Resilience is present in various functional units, for instance, in higher academic performance, emotional regulation, or communication competence [73].

McNeil-Willson et al. [119] researched Being resilient to violent extremism and polarization and their work addresses how to measure the risk factors for involvement in violent extremism. Resilience in this context is the opposite term to being prone to radicalization and engagement in violent extremism. Different studies that they referred to in their work give importance to the several factors to the building of social resilience to violent extremism, including; having positive community networks including a positive cultural image and identity, trust, and confidence in the members of the out-group as well as governmental institutions. In Shortland et al.'s study [165], they found that the exposure to extremist propaganda led to an increased pro-social behavior.

Social resilience is an inevitable excursion when speaking about polarization. It emphasizes humanity, pluralism, and values that hold together democratic societies. Social resilience within communities and large populations is an integral element in local and governmental disaster management and anti-terror programs that aim to give the individuals tools to cope with extreme distress.

### 4.5.3 Psychological Phenomena and Harm Following Exposure to Visual Violence

Exposure to extreme visual violence dissolves varied psychological reactions and phenomena that change from one individual to another based on multiple factors. These phenomena often serve malicious parties, and in particular extremist groups and terror

organizations, as part of their psychological warfare. The literature includes also the terms of media violence and graphic violence to describe visual violence. The indirect exposure to extreme violence through cyberspace can be a traumatic event in itself and the literature refers to it as a “secondary traumatic event”. Newer studies indicate that symptoms related to direct exposure to trauma can also be present in indirect exposure to trauma, also in the long term [103] [144]. These include desensitization, the cathartic effect, and the glorification of violence, and how collective memory and public opinion are a subject to manipulation by triggering moral outrage.

### Desensitization to Violence

Desensitization designates the decreasing of emotional reaction to the level of becoming numb, as Rabinovich [145] thoroughly discussed. Desensitization is per se not always negative, as it serves as a coping mechanism during overwhelming situations and helps in emotional regulation. Emotional regulation plays a healthy role in performing in a more optimally in increased stress situations, by adjusting the emotional state to the situation. Boyd and Swanson [23] mention that the exposure to virtual violence by those who do not have the tools to process it can lead to emotional desensitization. Mrug et al. [123], researched the emotional and physiological desensitization that results from real-life and movie violence in young adults acknowledging that this group is especially vulnerable to media violence, as the majority of the youth nowadays are exposed to violence from consumed media. Their research found out that desensitization is a short- and long-term physiological and emotional reaction to recurring encounters with violence and they claim that eventually it leads to the tolerance of violence and thus further violent behavior. Rabonovich [145] also found in her literature analysis that the majority out of forty viewed articles draw a correlation between emotional regulation, desensitization, and tolerance. Nevertheless, some of them could explicitly follow the connection between the progressive exposure to extreme violent content and the development of desensitization.

Therefore, desensitization has a higher impact on society from blocking empathy and becoming indifferent to violence, to tolerance and normalization of violence. Florea [56] also mentions the term “conformation”, which psycho-sociologists use to describe cases in which viewers change their attitude or behavior corresponding to the norms shown in the mass media. Hence, desensitization combined with conformation changes social norms regarding right and wrong behaviors which gradually transforms society into more extreme and violent.

### The Cathartic Effect

The opposite phenomenon to desensitization and emotional numbness is the cathartic effect whose meaning stems from ancient Greek and stands for the purification of emotions and releasement of accumulated tension<sup>10</sup>. Florea [56] researched in her work the connection between the exposure to visual violence consumed via media (mostly television) and

---

<sup>10</sup><https://www.merriam-webster.com/dictionary/catharsis>, Accessed on: 19.06.2024

the cathartic effect. She observes a clear correlation between the need for liberation of negative feelings, frustration, and aggression, which the catharsis achieves, and the source for those emotions in the exposure to media violence. Psychiatry observes the cathartic effect as a therapeutic one that helps release repressed emotions. The purification is expressed differently; for some, knowing that the aggressor was punished, brings them to the state of relief and completion. For others, as the psycho-analyst Sigmund Freud believed, the accumulated emotions and the following inner state of distress come back to peace once the aggressive act is done. Freud's theory hints at the potential escalation of violent behavior for viewers exposed to media violence.

This theory has been subjected to examination for years, especially in the context of violent video games, for instance by Robinson and Whittaker [153], and the findings of the different studies seem to be contradicting. Kühn et al. [97] point out in their study that their own finding of not having detrimental effects from playing violent video games stands in contrast with other experimental studies that did find resulted in short-term effects of aggressive thoughts and behavior. This question concerning the turning point of a person from either sensing feelings of aggression or holding radicalized ideologies to become an active aggressor is a main topic in the research of exposure to violence and there is no one deterministic answer to it [124]. We will not get any deeper into this discussion but acknowledge that individuals react differently to the exposure of violent visuals after which some have the urge to engage in further violence, and others release their negative feelings by being exposed to visual violence. The cathartic effect mechanism is strongly related to other phenomena, e.g. addiction to violence and the accompanied feelings, moral outrage, and glorification of violence.

### **Glorification of Violence**

The glorification of violence that we want to follow is the one associated with recruitment methods to terror organizations and their strategy to push people into engaging in extreme violent acts [10]. The glorification of violence is also in use in the arts industry, especially in movies (Tarantino's films are a distinct example), in the porn industry (especially the niche "snuff" which included rape and sexual murder scenes that are colored appealingly and promote violence [131]), or in the video games industry, which all go beyond our scope of research.

The glorification of daily- and terrorist- violence is a fringe reaction (that is socially accepted within the violent extremism group) to the sights of extreme violence practiced against the civil population and it also includes active martyrdom characterizes suicide attacks [150]. The glorification and mostly accompanied justification for violence exceed the level of a normative reaction to events that are clearly against all moral and human values and amplify the sense of adoration toward violence. Eventually, this gives extremist individuals the needed inspiration to commit similar violent actions and incite further terror attacks. The 2005 convention and EU directives even use the term "glorification-as-incitement" for the additional intent and causality components. The

visual component of those violent deeds is so transparent and accessible and is not based on complete imagination which gives another feasible dimension to the glorification [170].

### Moral Outrage

Moral outrage is the power that moves people, to speak up, to take action, and at times also to think irrationally. This gives extreme feelings to lead individuals. Moral emotions that are the affective responses to stimuli in different situations and contexts are responsible to moral outrage. This moral judgment is the key to determining good or bad intuitive emotions are the main driver [167]. The exposure to extreme graphic violence mostly involves the violation of morality. Grizzard [68] follows the exposure to graphic violence and how it serves as a moral motivator. The consumption of content from media strongly triggers the elicitation of strong emotions. Such graphic violent content can create such strong feelings of moral outrage, anger, and disgust that move individuals to be active, and lead them to the streets to protest, in the non-violent case. Often, these incited emotions make them feel at unease and seek out ways to change the situation while sensing high levels of frustration. Moral judgments are strongly connected to emotions, which are not impermanent and a subject to manipulation.

If parties with malicious intentions do it wisely, they can exploit the igniting of moral emotions to their benefit and draw people into action based on manipulation and their own interests. We have seen in 4.4.2, that terror organizations have facilitated their online infrastructure and social media is widely in use for their benefit and psychological manipulation. Moral outrage is a fundamental component in recruitment and the justification for violence. While political potential to exploit this phenomenon for populism and the design of public opinion is high, these parties are out of our scope, as we focus on extremist groups.

#### 4.5.4 Nationwide Mental Health Problems

The topic of mental health is strongly linked to psychology and stands at the core of psychiatry. Nationwide mental health problems are a rising huge concern for governments whose responsibility is to provide a safe environment, physical and virtual, through policies and legislation (e.g. anti-terror policies, child protection, etc.). The understanding that sustaining a healthy society in mind goes along with physical health as well and is crucial for a functioning society. Psychological terrorism is also one of the primary goals of terror organizations. Terror attacks worldwide result in major stress that statistically has a bigger impact on individuals on a national and international level than other horrors (for instance, car accidents), as Küey [96] explains. Harming a whole nation psychologically is a great tool for malicious parties and poses a great risk to society at large.

Levin et al. [103] and Levaot [101] highlighted that indirect exposure to mass trauma through consumption of news and media can have similar psychological effects usually referred to as direct exposure. The understanding that indirect exposure can be as well

harmful has been accomplished and more precautions are taken to prevent it, e.g. child protection on social media and traditional television<sup>11</sup> or social media platforms that have activated the blur filter on automatically recognized violent content and show a warning before viewers access the material<sup>12</sup>. Such trauma is referred to as *secondary trauma* and has varied impacts on individuals.

PTSD (Post-Traumatic Stress Disorder) is on the rise worldwide and more and more cases of exposure have developed from ASD (Acute Stress Disorder) which usually lasts a month from the trauma to PTSD that is diagnosed after at least one month. The trauma can be sexual assault, natural and man-made disasters, and combat. Individuals who suffer from PTSD are likely to have symptoms from four categories; intrusions (unwanted memories or flashbacks from the triggering event), avoidance (abstaining from trauma reminders), negative alternations in cognition and mood (including depression and anxiety<sup>13</sup>), and alternations in arousal and reactivity (becoming numb or having excessive arousal) [65]. On a nationwide, large populations that suffer from war or severe collective traumas, are often characterized by PTSD. This can have devastating consequences on a national level and it is within the governments' responsibility to take action, prepare and offer rehabilitation programs, and ensure a safe environment that would not make individuals re-live the trauma and reduce the chronicity of the post-trauma [120].

Many studies were investigating it, especially after the 9/11 attacks in the US. One of the many studies, which was presented by Neria et al. [128], found that the consequences of 9/11 on mental health exceeded the estimations and had short- and long-term consequences (e.g. significant functional impairment in the first four years following the attack). Another large study was done in another context by Levy-Belz et al. [102] on the Israeli population following the 7th of October attack, which belongs to one of the deadliest terror attacks in modern history. They found a significant increase in PTSD, depression, and GAD, for direct but also indirect exposure through media (we referred to in 4.5.1) and saw an increased risk for those with already existing psychiatric backgrounds. Al-zzwai [6] refers to child direct and indirect exposure to terrorist violence with explicit graphic content as an increased risk factor for PTSD, both in the short and long term. He mentions that graphic content can influence young individuals differently in the same population group but also from adults.

As the old Latin phrase states *Mens sana in corpore sano*, mental health has a direct impact on physical health. The connection between physical diseases following traumatic events and poor mental health has been repeatedly found. Neria et al. [128] found an increase in different diseases following the years of the 9/11 attacks, including diabetes, circulatory diseases, and lipid diseases Stress following exposure to violence

---

<sup>11</sup>Policy, Child Protection. "Child protection policy." Community Health 20.8250 (2021): 7333.

<sup>12</sup>Das, Anubrata, Brandon Dang, and Matthew Lease. "Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content." Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. Vol. 8. 2020.

<sup>13</sup>Denoted in the literature by GAD (General Anxiety Disorder)

involves also physiological mechanisms that lead in the long run to higher levels of blood pressure and cortisol that also have their share of poor physiological state, as hinted by Mrug [123]. Al-zzwai [6] also supports her findings in the context of constant exposure to graphic violence and states that these physiological changes have long term consequences on organ development.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Results

In this chapter, we integrate the chosen literature from the background with our processed analysis of our PRISMA corpus. Firstly, we define the primary risks of visual AI with the help of our suggested risks framework that utilizes several components that describe the risk level, the responsible actors, the contributing context, the first-level impacted group, and the harm analysis that refers to the nuances of the misuse of generative AI and potential outputs of graphic violence. We found eleven risks in our three upper categories. The category of radicalization has an important weight with seven found risks divided into three sub-categories. The categories of polarization consist of three risks and the psychological harm category two. In diagram 5.1, we introduce an overview of the categories and found risks.

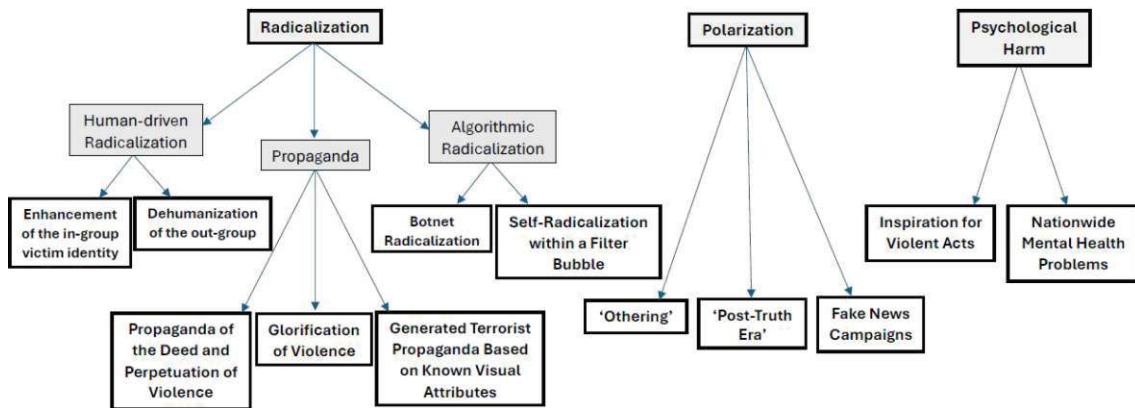


Figure 5.1: An Overview of the Found Risks

We then proceed with our gap analysis of the EU AI Act. We find that the act covers all relevant technological components, addresses many relevant actors along the technological chain, and its introduced risk levels largely include our risks by definition. However, we find

several gaps where we identify the need for future regulation and adaptations. Notably, the act lacks clear standards regarding admissible inputs and outputs, creating a substantial gap in managing offensive and graphic violence found in training data, potential user inputs, and harmful outputs. Our non-technology-specific gaps discuss the need to add malicious actors as possible stakeholders to the EU AI Act, as they are disregarded and the focus goes primarily to the restriction of misuse by public entities, and the unclear legal responsibility of media platforms in disseminating generated content, especially when the content is generated in another digital environment. The attempts to promote innovation through transparency might be harmful, as it supplies enough tools and information to malicious parties to operate and develop their own models. Furthermore, we cannot find any concrete reference to government-led education programs to enhance resilience by providing tools to consume generative AI content and use generative AI systems carefully and critically. Our following technology-specific recommendations emphasize the need for requirements related to human-in-the-loop moderation, the implementation of keywords and semantic analysis for prompts to proactively prevent the generation of extremist and violent content, and the integration of automated detection tools for outputs containing amplified violent content. We also highlight key points for future research and development of watermarking techniques as well as emphasize the need for visual anonymization through GAN models, and continuous user feedback integration, all offering solutions for nuances cases.

## 5.1 Primary Risks of Generative AI

Generative AI **text-to-image** and **text-to-video models** have great potential to contribute to society and parallel harm it on multiple levels in different settings, in particular in a violent context of wars and terrorism. The previous chapters have constructed a holistic overview of the technological background of Generative AI and its possible applications, the new type of violence in current times that includes live-streamed terror attacks in a “Big Brother” reality TV show style as a means of propaganda, the pathway of radicalization, how pluralism can turn into polarization, and the social and psychological (including psychiatric) consequences of exposure to extremely visual violence over the media. The media environment has become a particular contributor to the evolving risk landscape of generative AI, encompassing both digital and physical spheres.

The next step is to conclude, infer, and highlight the primary potential risks of employing this technology in times of war and terrorism and its role in exacerbating radicalization, polarization, and psychological harm. Therefore, we draw a line from the already acknowledged generative AI risks referred to in the background 2.3, the broad application of generative AI technology as a tool to generate manipulated visual outputs, and add the layer of our focus point of the current context, namely, having available extremely graphic violence and the concealed risks in exposure to such visual violence. The fundamental knowledge we acquired in our analysis helps us a step further to fully understand during what steps along the individual radicalization process and the structured radicalization

that utilizes propaganda, there is a potential hidden risk.

ISO 31000 defines a general risk as the “effect of uncertainty on objectives”<sup>1</sup>. It treats an effect as a deviation from expected outcomes and defines objectives as a broad term for different sectors applied at various levels. We work with this definition and focus on effects that have the potential to cause harm. Our own proposed framework defines each risk by a composition of several components that give altogether a comprehensive understanding of the risk. The description of the risks is nuanced and based on the literature introduced in the background combined with our acquired knowledge from our working PRISMA corpus. The several components are introduced in table 5.1.

Component	Meaning
<b>Risk level</b>	following the risk-based approach of the EU AI Act for risk management, we use four levels of risks: unacceptable risk, high risk, limited risk, and minimal risk. We follow the suggested definitions in the EU AI Act and also refer to the plausibility of the risk combined with the harm it can cause. Another leading factor in assigning a priority is the question of whether the named risk is the major reason for the harm or whether there are further external components that also influence the large phenomenon.
<b>Responsible actors for the risk</b>	includes the direct actors that cause the risk, either intentionally or unintentionally. When considering multiple potential actors, we prioritize high-risk actors based on their recent history of being responsible for mortality. Lower-risk actors include political groups within the democratic space that generally pose a lower risk, due to less mortality and the design of democratic regimes that should have protection mechanisms. Some may have the potential to undermine democratic principles and institutions, potentially paving the way for a shift towards autocracy.
<b>Direct / indirect harm</b>	the direct involvement of human actors in the cause for the risk.
<b>The context within the risk takes place</b>	the pre-conditions that create the context in which the risk holds. This includes the causing factors.
<b>The harmed target group</b>	this distinction examines how generally applicable the risk is or if it rather targets specific groups as well as observes the interplay between harming individuals and its implication to society at large. We mostly focus on the first-level impacted group.
<b>Harm analysis</b>	an assumption about the damage caused as a consequence of the risk to the target group.

Table 5.1: The Components of our Risks Framework

<sup>1</sup><https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en>, accessed on: 16.08.2024

To deeply understand the potential risk, the scenario of **generating visual content with violent characteristic**, each identified risk comprises three stages; the training, the immediate output with its corresponding short-term effect, and the long-term effect. In each stage, we will identify the potential for malicious parties to exercise and infiltrate manipulation and the possible **intentional or unintentional harm** that can be done. The risks include the technological capabilities to achieve specific content, which mainly belong to the first stage, and the socio-psychological aspects that rather belong to the two latter stages.

Generative AI technology is a compound of data and models. Currently, the big IT companies mostly provide it for commercial reasons. We can also base our risk analysis on the current state-of-the-art that has already reached the technological ability to generate visually violent content, as we have seen in the latest work of Hao et al. [72] from February 2024. In this work, we consider the worst-case scenario, namely, the models can get out of the hands of these companies and private parties can gain access to them and utilize both the data and/or the models. Hence, they would be able to operate their own models while bypassing the big IT companies' policies. We have seen a similar situation with the parallel private social media platforms used and operated by terrorist organizations. We suggest that these parties can be terror organizations, political parties, and even governments.

The need for cloud services is filled and dominated by a small number of huge IT groups, mainly Microsoft, Amazon, Google (Alphabet), and further Nvidia, Oracle, and IBM [176]. Most of the main Generative AI models whose risks we investigate also belong to those companies or have tight relations with them. For example, OpenAI's partnership with Microsoft, that during the time of writing Microsoft owns 49% of the company [176].

Having said that, the feasibility of the outlined risks might be questioned, but we would rather include also the risks that technologically might not be fully feasible at present. One of our objectives is to raise awareness about the potential manipulation inherent in generated visuals and their direct purpose. Critical thinking is a vital tool, particularly as we move forward into the future. As Generative AI models rapidly evolve, regulations struggle to keep pace and global (political) interests shift, obtaining a comprehensive understanding of the future landscape is critical.

### 5.1.1 Radicalization-related Risks

The structure of terrorist organizations is complex networks that advocate violence to achieve some goals that are part of an extremist ideology. These ideologies are mostly against Western values and want to change society and its norms. The path to terrorism goes through the radicalization of individuals that would eventually lure them into joining the organization. The risks of generative AI that belong to this category include human-driven and algorithmic-driven radicalization and advanced means of propaganda. This category has an important weight in our risk landscape, due to its high misuse

potential and diverse applications along the radicalization processes. Having acquired the foundations for radicalization processes from the working corpus, we identify mostly the potential of malicious actors to cause direct harm, utilizing this technology to design advanced propaganda campaigns on a personalized base in high volume.

### Human-Driven Radicalization

Human-driven radicalization includes radicalization models that are human-designed and in human use. In 4.2.3, we viewed two radicalization models, namely, Webber and Kruglanski's [94] 3N, and Sageman's [160] four-staged radicalization models. Both models share the core attributes:

- The initial discontent of individuals who seek a change to handle their feelings of victimization or grievance
- the moral outrage that amplifies the “othering” mentality, and the feeling of belonging to some group with strong identity feelings.
- The group dynamics that by structure repeatedly confirm the initial feelings of discrimination and bias and lay the foundation for radicalization.

Generative AI models enable the quick and smooth generation of different visuals that can contribute to these components. Lakomy [98] spoke about the risk of using generative AI to generate content that fits in terrorist campaigns. Puczyńska et al. [143] also mentioned that generative AI can contribute to radicalization via the creation and distribution of disinformation. Esmailzadeh [53] adds the layer of portraying the enemy in a bad light with the help of generative AI. Our following presented risks draw a line between these known risks to the components of the radicalization models, and in particular dehumanization.

**Enhancement of the in-group victim identity for terrorist causes.** This **direct risk** involves human actors. These higher-risk **actors** primarily include the violent extremist groups and terror organizations that want to recruit and benefit from the pre-existing discontent of individuals and enhance the feelings of inferiority. They pose an **unacceptable risk** level to society, due to their terrorist nature and misuse potential of deploying subliminal techniques to manipulate individuals. The manipulation aims to move individuals into actively joining violent organizations that pose a great risk to society. The specified use includes the enhancement of the “othering” phenomenon while focusing on the in-group identity while utilizing means of *dehumanization* against the out-group members, accusing them of the suffering of the in-group. The radicalization pathway gives **the context** for this risk, including the above-mentioned components of the radicalization models. The stages of influence for the potential manipulation include the following steps:

- I. Training data that detects characteristics of victims of violence, could be enhanced with ethnic characteristics. In combination with the understanding of the aesthetics of violence
- II. Immediate output triggers a moral outrage due to the hyperrealistic demonstration of the suffering of the in-group
- III. This moral outrage is seen long-term in the design of individuals' opinions and the public opinion within the in-group and accumulating feelings

The enhancement of the “us vs. them” by putting focus on the in-group narrative of being a victim is the main consequence of this risk. The utilization of Generative AI systems to generate specified output for personally enhancing these feelings takes benefit of known group's characteristics and historical events that shape public opinion of its members. The tool of generative AI for exploiting these feelings can manipulate individuals within a specific group that sequentially obtain a confirmation of their feelings of discontent. Their feelings are backed up by a visual justification for the 'rational' fear from the out-group. This risk applies both to a specific target group and to the general society:

- I. Generally, individuals can get lured into joining terror organizations and engage in violent acts in the worst-case scenario. As we explained in 4.2.2, individuals do not necessarily have to be prone to radicalization. Nevertheless, individuals who already feel discriminated against or victimized can be subject to personalized radicalization, based on their characteristics and identity. Structured groups can develop an “automatic personally designed manipulation”.
- II. Individuals who belong to the out-group, who are allegedly responsible for the suffering of the in-group, are the second target group of this risk. They suffer from a bad reputation that is strongly influenced by the demonstrated violence they have engaged in.
- III. There is an implicit general risk to a society suffering from increased violent acts and terrorism, as more individuals will eventually engage in violence through these personalized recruitment methods.

This risk enhances the feelings of victimization through generated hyperrealistic images and videos, that malicious parties can personally adapt to target individuals of a specific group and reach the mass crowds that might feel identified. Parallel, the out-group is portrayed viciously, due to its responsibility for the caused suffer. We call it the first type of dehumanization, as the out-group takes inhumane acts.

**Dehumanization of the out-group.** This **direct risk** also involves human actors. These higher-risk **actors** primarily include the violent extremist groups and terror organizations that want to enhance the sense of inferiority toward the out-group members. In this case, they also pose an **unacceptable risk** level to society, due to their terrorist nature of encouraging individuals to engage in violence toward a

targeted group with the help of subliminal techniques. The manipulation in this specified use aims to normalize violence targeting the out-group members by assigning them inhumane attributes. We have mentioned in 4.3.2 the Nazi propaganda against the Jews that used a similar tactic toward this ethnic group. The normalization of violence toward a specific group eventually triggers a less emotional reaction to such cases. This phenomenon also describes the desensitization to violence that we discussed in 4.5.3. An extreme example of dehumanization is the celebration of bodies by members of terror organizations. The pre-established radicalization pathway gives **the context** for this risk, including the above-mentioned components of the radicalization models. This type of dehumanization stems from the group dynamics that navigate the feelings of moral outrage and in-group identity to action against the out-group. Violent extremists are cognitively intolerant to the out-group members and dispose of moral duties towards them. Terror organizations take it to the next level and actively endorse organized physically violent acts. The stages of influence for the potential manipulation include the following steps:

- I. Training data that detects characteristics of the out-group together with the understanding of the aesthetics of violence
- II. Immediate output that assigns hyperrealistic inhumane attributes that often go along with violent attributes
- III. This normalization of violence contributes in the long term to the desensitization of violence toward a specific group while designing the public opinion to omit moral duties toward it

As part of the dehumanization that often takes place within the framework of “othering”, mass graphic violence toward the out-group that AI generates can emphasize its inferiority, which can cause great harm. In extreme cases, if done wisely, showing out-group members who suffer from extreme violence, especially with the various hyperrealistic outputs of the generative AI models, might cause targeted feelings of dehumanization, which also go along with the development of desensitization and the decreasing of empathy toward the out-group members. This risk is mainly applicable to the members of the specific targeted group who may suffer from exercised violence and less empathy in times of need from non-group members. Therefore, they will be more vulnerable to external risks.

This risk enhances the feelings of inferiority of the out-group vs. the superiority of the in-group through hyperrealistic exercised violence toward the out-group. The resulting annulling of moral duties toward the out-group gives the legitimization to commit violent acts against it and normalizes them through desensitization.

### Propaganda

All of the papers discussed in the 2.3 that highlight the terrorist use of Generative AI, mention propaganda purposes as a main threat that aims to recruit individuals. We

have seen that propaganda primarily promotes extreme ideologies that are often binary and simple to understand to swap away the mass crowds. There are different ways to convey propaganda and the visual style, aesthetics, and the appeal of the content play a significant role in the dissemination methods.

**Propaganda of the Deed and Perpetuation of Violence.** This is a **direct harm** that involves human actors who have carefully designed these visual aesthetics. These higher-risk **actors** include mostly terror organizations that see violence as a main part of their communication strategy and signature and means to convey their message. We would like to mention a secondary risk actor that is not our focus point, namely totalitarian regimes. We have discussed this actor briefly through the multiple examples of the Third Reich's tactics. Totalitarian regimes employ terrorizing methods, such as public executions. The use of such violence can also relate to a sort of propaganda of the deed. This risk poses an **unacceptable risk** level to society, due to their terrorist purpose to promote acts of extreme violence while deploying subliminal techniques that exceed the individual's consciousness. The rising recent trend of the propaganda of the deed, which promotes the acts of violence and bloody images as the propaganda itself, creates the **context** of this risk. Ferrera [54] refers to the risk of generating offensive content as a risk by malicious parties. Lakomy [98] did not manage to generate graphic content of dead people in his study and mentioned it was due to anti-terror mechanisms that the regulation enforced. Technically, it is possible to reach graphic bloody content, as Srinivasan et al. [72] found in their research. The stages of influence for the potential manipulation include the following steps:

- I. Unfiltered training data can easily cause such scenarios for IT companies' owned models. Under the hands of terror organizations, models can be completely trained with their specified content to serve their terrorist needs.
- II. The output can be used as a tool to promote the propaganda of the deed and to misuse graphic violence as part of targeted psychological warfare.
- III. The long-term effect is drawn from the way the trauma is processed by each individual and is therefore not unified, as we have seen in 4.5.4.

As part of this risk, malicious parties can misuse multiple technological features of generative AI models:

- I. The task of generating photos or videos that are extremely bloody and include graphically hard scenes to watch. Generative AI models do not have human sensitivity or mental capacity when exposed to extremely hard content, and without programmed or trained identification that the content went too far, for the models these are just further outputs, and for the variation of violence, the sky is the limit

- II. The task of bringing to life static images<sup>2</sup>, which some of the generative AI visual models enable. This task could further be applied to images that are clearly graphically violent, for example, a photo of injured or dead people, and bring the violence to life, letting the Generative AI model portray a whole imaginary horror scenario. This could be done even with a censored photo, as further advanced AI features enable fixing of blurred or pixel photos<sup>3</sup>. The ability to 'bring to life' the photo into a complete video that fills the gaps based on the pre-given violent attributes can lead to hallucinated horror movies of violent scenes
- III. When it comes to targeted campaigns, malicious parties can use photos from real-life occurrences and generate scary visuals to terrorize on personal levels individuals. We identify the potential to generally cause harm to individuals who are the subject of a blurred image or video, as their anonymity might no longer persist, which also creates a privacy problem. This could cause additional pain and trauma to the victims and their families.

This risk applies both to targeted individuals, as explained, and to the general society that has a higher risk of coming offensive content. This risk perpetuates violence toward humans and serves a varied range of malicious parties as part of their terrorizing techniques.

**Generated Terrorist Propaganda Based on Known Visual Attributes.** This is a **direct harm** that involves human actors who base their communication strategy on visual aesthetics. These higher-risk **actors** include mostly terror organizations with distinct visual characteristics and the secondary risk actors of totalitarian regimes that are not our focus point. This risk poses an **unacceptable risk** level to society because it greatly helps facilitate terrorist infrastructure that aims to affect individuals and groups and their behavior. Lakomy [98] managed in his study to reach generated outputs of a combat style that terror organizations can adapt to their needs, but could not reach results from prompts with obvious terrorist terms. Furthermore, we have discussed in 4.5.1 the distinct visual characteristics of both ISIS with their beheading videos using their repeating logo and dramatic music, and the Nazi regime with its distinct visual signature that is part of their communication strategy. These unique attributes serve as the **context** for this risk, which can very easily be translated into training tasks for generative AI systems. The clear association between visual attributes and a terror organization or political parties is also part of the propaganda. The stages of influence for the potential manipulation include the following steps:

<sup>2</sup>An example for making images come to live has already been demonstrated by Google to the public makes Mona Lisa alive, <https://www.youtube.com/watch?v=P2uZF-5F1wI>, access on: 28.06.2024.

<sup>3</sup>[https://store.google.com/intl/en\\_uk/ideas/fixed-on-pixel/](https://store.google.com/intl/en_uk/ideas/fixed-on-pixel/), accessed on: 28.06.2024.

- I. The training phase that puts focus on recognizing visual attributes and their belonging and the relevant training data including the relevant content stand at the core of this risk
- II. The immediate output is here a trivial risk in which propaganda material can be easily generated based on the known visual attributes
- III. The long-term effect involves the endless pool of propaganda material and the following recruitment or gained motivation within the terror organization

It is highly plausible that this **universally applicable** risk would soon generate AI-powered propaganda that is easily and quickly created, both conceptually and visually, either for old or new ideologies and organizations. This propaganda is extremely dangerous, as it can be highly persuasive thanks to training methods that utilize data that is known to have worked in the past. Generative AI models contribute to these two attributes and push us to a whole new era of personal yet automated content that looks hyper-realistic.

**Glorification of Violence.** This is inherently an **indirect harm** that the technological capability of generative AI models to generate outputs that demonstrate violence in a glorified way and make it look hyperrealistic causes. Terror organizations are **higher-risk actors** that can exploit this technology to glorify their acts of violence, as part of their communication strategy. Social media platforms that promote the virality of such images are secondary actors, as it should its within their responsibility to inform the user about the origins of the content if it is not human-generated. This risk poses a **limited-risk** level because it has the potential to manipulate individuals and the artificial origins of the output should be transparent to all. The **context** of this risk is due to the fact that the glorification of violence is a major component of the propaganda of the deed and is very common in recent terror organization communication strategies, in times taking its inspiration from video games, as we discussed in 4.5.3. The stages of influence for the potential manipulation include the following steps:

- I. The training data is crucial in this case. A specific manner to manipulate the data is the use of video games in the training data in which violence is demonstrated in an aesthetic way
- II. Immediate output aims to demonstrate violence appealingly to lure individuals
- III. In the long term, it could result in an increase in terror attacks designed to maximize visual impact. By equating generated graphics, real-world violence that produces bloody images would be promoted and glorified

This risk is **universally applicable** and stems from the glorification ability that generative AI excels at. Visually the outputs from generative AI models are impressive and have a certain aesthetics due to the complex algorithms and techniques that create each visual, as we discussed in 2.1. Terror organizations exploit it and misuse methods of glorification when to promote their violent deeds.

It helps them in recruiting and motivating people to engage in terrorism, like Puczyńska et al. [143] example of suicide terrorists that have a happy face and hide the actual physical suffering that goes along with an attack. The exploitation potential of such systems is clear and worrying. The wisely manipulated content can drive people to act violently and even get inspiration from such systems.

### Algorithmic Radicalization

Algorithmic radicalization includes technological infrastructure that aims to facilitate cyberspace in a personalized manner and bring individuals with similar characteristics and interests together. This finally leads to segmented networks over the internet, and particularly over social media platforms, in which individuals are exposed to the same suggested content. Throughout time, the preferences get enhanced through feedback loops and the digital identity of the users. Puczyńska et al. [143] mentioned the risk of radicalization through personalized information bubbles together with the application of social engineering methods. Algorithmic radicalization is a compound of societal radicalization processes and disseminated propaganda and could involve human intervention with botnet campaigns or just evolve through the algorithmic structure of social media platforms and their ability to learn about the characteristics of their users, particularly thanks to centralized identity providers.

**Botnet Radicalization.** This is a **direct harm** that involves human actors who utilize sophisticated methods to disseminate propaganda. These higher-risk **actors** include terror organizations and violent extremist groups that aim to both terrorize mass crowds and recruit individuals. Social media platforms that enable such malicious campaigns and do not make any efforts to identify and stop them are secondary actors, especially because they provide the infrastructure for such attacks. This risk poses an **unacceptable risk** level to society and infrastructure, due to their ability to take over virtual spaces in DoS-like cyber attacks or reach individuals with precision based on their characteristics. Such a bot activity can increase the exposure of society at large to offensive content on the one hand, or target prone individuals on the other hand. The technological capability to quickly generate different outputs yet with similar visual style provides the **context** of this risk. We mentioned in 4.5.1 Al-Khateeb and Agarwal's work [4]. They explained how the violence that is evident in the content from ISIS was a significant contributor to its virality and that the network of bots was largely responsible for its successful dissemination across the virtual space. The stages of influence for the potential manipulation include the following steps:

- I. Training models to output vast volumes of creative outputs that have similar visual characteristics. If we consider the anti-terror mechanisms that social media platforms should integrate, a large number of outputs of generative AI models can be designed by malicious parties to automatically observe the limits of censorship of these platforms.

- II. The immediate output can promote violent propaganda or misuse the graphic content for targeted psychological warfare.
- III. The long-term effect is recruiting individuals who have engaged with radical content online or cause trauma to large populations in society. The way individuals process trauma is individual and is therefore not unified, as we have seen in 4.5.4.

The risk applies to the following target groups:

- I. Generally applicable to society as a whole who might suffer from increased offensive content packed in a cyber attack over social media infrastructure
- II. Individuals who are individually approached by terror organizations during the “love bombing” phase, as we discussed in 4.4.2. This risk strongly connects organized terrorism and cross-technological advances that can be easily exploited by these malicious parties to promote their agenda and organizational needs

**Self-Radicalization within a Filter Bubble.** This is an **indirect harm** that the inherent structure of social media platforms creates. The higher-risk **actors** that benefit from this technological tendency are extremist groups and terror organizations that effortlessly manage to reach individuals based on their digital identity and preferences. The personalized internet infrastructure and social media platforms also contribute to this risk as a secondary risk actor. This risk poses a **limited-risk** level to society and infrastructure, due to its potential to manipulate individuals into holding radicalized or polarized opinions and repeatedly receive confirmation from their peers in the same filter bubble. The technological infrastructure together with the fundamentals of radicalization models provide the **context** to this risk. It enhances the “othering” phenomenon and exposes individuals to topics that can trigger moral outrage, also if they are physically distant. The generated content can only amplify these feelings, based on known characteristics of these individuals. The stages of influence for the potential manipulation include the following steps:

- I. Training models that aim to integrate in their outputs visual characteristics that particularly influence emotional reactions, based on known patterns
- II. Immediate output triggers a moral outrage
- III. The long-term effect is expressed in the design of public opinion through manipulation of the accumulating feelings of individuals

Radicalized opinions can get more prevalent, as individuals shape their opinions based on hyperrealistic graphics that disregard any critical thinking by having a highly convincing appearance and not being exposed to opposite opinions or world views. This could also incite single violent actions of individuals who find significance and inspiration in radicalized topics and such content. The risk applies to the following target groups:

- I. Individuals who would be motivated to engage in violent acts and join terror organizations in the worst-case scenario. This risk is generally relevant, as we have seen that everyone is prone to terrorism. This risk is the extension of recruitment strategies and encourages the terrorist presence on social media platforms
- II. There is an implicit general risk to a society suffering from increased violent acts and terrorism, as more individuals will eventually engage in violence through these personalized recruitment methods

This risk is inevitable, due to the complex algorithmic structure of social media platforms. As long as there is an extremist or terrorist presence on these platforms, their emotionally triggering content will reach individuals in a personalized manner.

### 5.1.2 Polarization-related Risks

Several experts have raised their concerns about the use of Generative AI for political polarization. Puczyńska [143] briefly spoke about the polarization risk through personalized filter bubbles. WEF’s report [58] strongly connects the risk of misinformation and disinformation to societal polarization and refers to it as one of the main risks for 2024.

“**Othring**”. This is a **direct harm** that involves human actors who utilize sophisticated manipulation methods to disseminate political messages. The large emphasis on the differences between the in-group members to the out-group members while shifting ideologies to their extreme gives the **context** of this risk. Parties on the extreme political range are the **actors** that benefit the most out of this great division within society, as part of the phenomenon of the “disappearing middle” which we have viewed in 4.5.2. This risk is of a **limited-risk** level, as it provides additional tools for manipulation that amplifies an existing societal phenomenon. The lack of transparency regarding the origin of the disseminated information is key to regulating this risk. The stages of influence for the potential manipulation include the following steps:

- I. Utilizing training data sets that align with fringe ideas creates a rooted bias toward these worldviews
- II. The immediate effect involves the moral outrage that plays a significant role in manipulation and accelerates societal division
- III. The long-term effect is the major concern of this point, as it brings about shifts within society and emphasizes the inter-correlation between technology and society

Manipulated AI violent outputs can enhance the process of ‘othering’ in the context of polarization. This eventually leads individuals to believe that there is no common base with the out-group members and have enhanced feelings of identification with the group of belonging, based on the common suffering of the in-group. Extreme

political parties can wisely employ AI technology to generate content that would benefit their own political needs. Polarized societies are characterized by decreasing empathy and civil courage and the additional increasing mistrust in the democratic institutions or sources of information that are associated with the out-group. Manipulated visual content, particularly graphic violence can play in the favor of extreme parties that can easily base their fringe ideas on manipulated visual justification. The risk applies to the following target groups:

- I. Targeted groups by extreme parties who can utilize pre-knowledge to manipulate whole communities.
- II. The risk is eventually universally applicable, due to the final outcome of a polarized society.

This risk reflects a societal tendency that can only be enhanced through the capabilities of generative AI to illustrate a manipulated reality.

**“Post-Truth Era”.** This is an **indirect harm** that stems from the combination of decentralized systems and the flow of information. The post-truth era, which we discussed in 4.5.1, and the increasing confusion about facts, from either historical or present occurrences, provide the **context** of this **limited-risk**. The involved **actors** are the platforms for news distribution, both traditional and social media platforms. The increasing anonymity online, which we viewed in 4.3.2, enhances the distribution of irresponsible and fake information bringing along last years’ phenomenon of fake news. Extreme political parties, extremist groups, and terror organizations benefit from this phenomenon to promote their ideology through the increased confusion that dominates society. Together with the decentralization of news resources and information channels that has led to the “everyone is a journalist” phenomenon, pieces of information are trusted regardless of their credibility. Instead, they are mostly based on the individual’s emotions. This post-truth era is already widely present in the last couple of years and due to the large supervised (e.g. mainstream social media platforms) and unsupervised communication channels (e.g. telegram and private social media platforms), it is easier than ever to distribute fake content. The virality of such content is largely determined by how sensational and visually attractive is the piece of information.

The large information campaigns coming from different directions online that include violent content only amplify this ongoing phenomenon. The exposure to brutalities happening during times of war and terrorism from non-traditional media, which traditional media would most probably censor strictly, lures viewers to consume their information mainly from such channels and disregard the credibility of traditional media or even official information that governments or international organizations publish. Obtaining information from different channels gives on the one hand the feeling of control and of understanding the whole picture. On the other hand, it can lead to feelings of confusion, (mis)trust in nothing and everything,

and being lost in the mass information. Especially, in physically remote situations, the only relation to them is second-hand obtained information that is delivered first-hand through visual images and videos supplying an immersive feeling. The stages of influence for the potential manipulation include the following steps:

- I. Systems with no measures of identifying generated extremely violent graphic content can as a result output any arbitrary violent content that can be used by any malicious party with interests. For example, no watermarks.
- II. The immediate effect is the overload of information and increasing confusion and tendencies to believe news and facts based on emotional reactions and not on their credibility.
- III. long-term effect is the post-truth era and the strong mistrust in credible pieces of information or trust in information based on arbitrary emotions to the graphic content.

The hyperrealistic, graphically violent content that AI models can generate enhances mistrust and misdirects it toward unreliable sources. Thus, it deepens widespread confusion among large populations. The risk is, therefore, **universally applicable** and leads to a general mistrust in information.

**Fake News Campaigns.** This is a **direct harm** that involves human actors. The **context** of the post-truth era allows malicious **actors** to take advantage of the general mistrust that already dominates society to disseminate manipulated information. This is a **high-risk**, as it has the potential to lead to complete societal chaos when misinformation and disinformation get out of control. The stages of influence for the potential manipulation include the following steps:

- I. Systems with no measures of identifying generated extremely violent graphic content can as a result output any arbitrary violent content that can be used by any malicious party with interests. For example, no watermarks can be a great contributor to large fake news campaigns
- II. The immediate effect is the overload of manipulated information and increasing confusion among the population.
- III. The long-term design of public opinion and reaching the point of no return in which also hard facts will not be trusted because the truth would become questioned.

Generative AI models facilitate and ease large campaigns for distributing misinformation. It is by now also technologically feasible to fabricate past events and rewrite narratives in history with “visual facts”. Having hyperrealistic visuals from historical events that are unidentifiable whether they are real or fake can play as a great tool of deception and is, therefore, **universally applicable**. George Orwell wrote in his book 1984 that ‘who controls the past controls the future: who controls the present controls the past’. This phrase could not have been more relevant to our scenario, especially as part of the long-term risks.

### 5.1.3 Psychological-related Risks

Exposure to visual violence and radicalized content can have significant psychological impacts on individuals, often varying based on different factors. These effects range from positive positive, such as fostering resilience in individuals, communities, or nations, to detrimental, both individually and collectively. While malicious actors may not be solely responsible for these effects, they can benefit from the psychological harm that often results. In the context of exposure to graphic violence, potential negative consequences include increased violent tendencies in society and inspiration for engaging in violent acts, and widespread mental health problems.

**Inspiration for Violent Acts.** This risk causes an **indirect harm**, as it is mostly technology-driven and reflects through new variations the given set of training data. We identified this risk through a combination of the propaganda generator that we have thoroughly discussed in the earlier risks, individuals who are inspired to engage in violence based on visual violent scenarios they have been exposed to, and psychological phenomena of the cathartic effect and glorification of violence. It poses a **limited-risk** level, as this type of tool is not the only source of inspiration for violent acts if we consider other violent movies, books, or video games. Yet, due to the hyperrealistic visual outputs, it is necessary to disclose their artificial or manipulated origin. The relevant **actors** that can exploit these risks are first and foremost terror organizations. We also include other radical and violent extremist groups as a possible actor. Lastly, we include individuals who are prone to engage in violence, like lone wolf extremists, who find their significance by exercising violent acts. The increasing volumes of accessible glorified violent content provide the **context** together with the range of creative and hyperrealistic scenarios that the generative AI models can produce. The stages of influence for the potential manipulation include the following steps:

- I. There is no filter for extremely violent content during the training phase which can lead to highly violent outputs. The systems themselves might also not have any checkpoints to ensure that the output conforms with the security of the users. This happens especially when there are no standards to decide whether the output is conforming or not. During the training phase, the systems learn the different possibilities for exercised violence and can perpetuate these possibilities in the future outputs, based on the user input or other model hallucinations
- II. The immediate exposure to the violent output creates acute stress for some individuals. Others experience a cathartic effect and might be motivated or inspired to repeat the scenario in the physical realm
- III. In the long-term, outputs of graphic violence from generative AI models can trigger trauma for individuals, also even through a one-time exposure. Another possible long-term consequence is the normalization of violence that triggers a less emotional reaction over time.

Generative AI models can potentially create an unlimited pool of new visual violent content and be treated as an 'inspiration generator' for individuals or organizations. This is relevant to individuals with a rather twisted mind but could also pose a risk to normative individuals who are prone to fall for extremism or those who suffer from violent tendencies. The combination of a lack of moral values and the appealing hyper-realistic images and videos that these models can generate has the potential to amplify dangerous and fringe tendencies that already exist in society and bring new ideas about how to engage in violent acts, and is, therefore, **universally applicable**.

**Nationwide Mental Health Problems.** This is inherently an **indirect harm** that usually accompanies extreme stress situations and triggers different psychological reactions individually. Terror organizations are **higher-risk actors** that can exploit this mechanism following their attacks to proceed with a secondary aim of psychologically harming large populations. This risk poses a **high-risk** because it has the potential to harm the functionality of society on a large scale and has further implications in the short- and long-term. The mental strain following the media exposure to extreme violence creates the **context** of this risk. Terrorism and exposure to bloody visuals can lead to traumatic symptoms similar to those who have experienced direct exposure. The high-risk **actors** that are involved in this risk are trivially terror organizations that launch attacks and receive broad media attention to it, both on traditional media and social media platforms. Another significant actor is governments whose role is to prepare resilience programs that would help minimize the damage and the vulnerability of society. Part of the resilience programs include national programs to deal with traumas, and education programs for careful and responsible use and consumption of generative AI. Malicious parties that manage to get access to generative AI systems or exploit publicly accessible models would have the ability to create huge damage to the public health of many individuals, which is the imputed liability of governments. The stages of influence for the potential manipulation include the following steps:

- I. If there is no filtering against extremely violent content in the training phase, it enables the output of further generated violence scenes. Systems can be exploited if they do not have any checkpoints to ensure that the output is secure (or "user-friendly"). Another potential to misuse the systems is the lack of official standards to determine whether an output is conforming or not
- II. The immediate exposure to the violent output creates an acute stress
- III. The real danger of graphic violent outputs from generative AI models is the long-term effects that can be triggered by even a single exposure

This **universally applicable** risk is a consequence of amplified violence by generative AI models that alone can enhance the already existing risks of different mental health diseases, e.g. PTSD, depression, anxiety, and ancillary physical health problems. There is a gentle interplay between the resilience of a society

## 5. RESULTS

that gets especially developed during hard times and its mental destruction during these times. Mental health is an extremely complex topic and individuals react differently to the same situations of exposure, as we discussed in 4.5.4.

Our results include ten primary risks associated with the use of visual generative AI models in the context of terrorism and war within three different categories: radicalization, polarization, and psychological harm. While other authors have briefly touched on these risks viewed in 2.3, our analysis provides a comprehensive exploration of each, revealing the full scope of the risk landscape. The key findings are highlighted in the following table 5.2:

Category	Risk	Misuse of Generative AI	Risk-Level	Key Actors	Direct / Indirect Harm
Radicalization	Enhancement of the in-group victim identity for terrorist causes	Generating outputs that enhance the feelings of the in-group's victim identity, exploiting known group's characteristics, and shaping historical events. Demonstrating a visual justification for the negative feelings toward the out-group that is portrayed as the cause of the suffering of the in-group.	Unacceptable	Terror organizations, violent extremist groups	Direct
	Dehumanization	Assigning with generative AI visual inhumane attributes to the out-group, including mass graphic violence exercised against it. Eventually, developing desensitization to violence toward the out-group.	Unacceptable	Terror organizations, violent extremist groups	Direct
	Propaganda of the deed and perpetuation of violence	Multiple tasks that generative AI models can perform include varied outputs of extremely violent scenes, transforming hard images into videos that display generated violent scenes, or unblur scenes from real-life events that aim to terrorize individuals.	Unacceptable	Terror organizations	Direct
	Generated terrorist propaganda based on known visual attributes	Generating vast amounts of AI-powered propaganda quickly and easily, utilizing concepts that the data shows to have been working in the past.	Unacceptable	Terror organizations, violent extremist groups	Direct
	Glorification of violence	Generative AI models excel at hyperrealistic outputs with outstanding esthetics which largely contributes to the glorification of violence phenomenon.	Limited	Terror organizations	Indirect
	Botnet radicalization	The vast amounts of similar content that generative AI can produce can be used in a botnet attack similar to the concept of DoS-attacks.	Unacceptable	Terror organizations, violent extremist groups, social media platforms	Direct
	Self-radicalization within a filter bubble	The vast amounts of similar content that generative AI can produce and its high distribution can create filter bubbles of violent content.	Limited	Terror organizations, violent extremist groups, social media platforms	Indirect

Figure 5.2: Terrorist and Extremist Risks of Misuse of Generative AI

Category	Risk	Misuse of Generative AI	Risk-Level	Key Actors	Direct / Indirect Harm
Polarization	'Othering'	Generating graphic violence that can play in favor of extreme parties to increase the general mistrust in democratic institutions for their own political needs.	Limited	Extreme political parties	Direct
	'Post-truth era'	Increased high flow of information from decentralized sources by easily generating vast amounts of content with generative AI models together with integrated violence enhances the emotions and moral outrage and sows more confusion about the facts.	Limited	Platforms for news distribution, extreme political parties, extremist groups, terror organizations	Indirect
	Fake news campaigns	Facilitating with generative AI large campaigns for distributing misinformation and manipulating historical facts with so-called visual proofs.	High-risk	Any malicious actor	Direct
Psychological Harm	Inspiration for violent acts	Creating with generative AI models an unlimited pool of diverse violent acts inspired by past violent acts that could be an 'inspiration generator'.	Limited	Terror organizations, violent extremist groups	Indirect
	Nationwide mental health problems	Generative AI models do not have the same moral sense humans have and are not aware of the psychological limits of violence and offensive content that individuals can be exposed to. Violent outputs can therefore enhance already existing risks of mental health problems.	High-risk	Terror organizations	Indirect

Figure 5.3: Terrorist and Extremist Risks of Misuse of Generative AI (continued)

Human actors are primarily responsible for the intentional misuse of these models. We have categorized these risks as limited-, high-, and unacceptable-level risks, aligning with the risk definition and risk-based approach of the EU AI Act.

We have presented the most relevant risks of generative AI systems with a great misuse potential and deviation from the original aim. The primary malicious actors are terror organizations and extremist groups. We see that most of the risks are universally applicable which portrays how significant these risks are to society at large. Due to the generality of generative AI models with a visual output that are the core component of many AI systems, the mitigation of these risks is not trivial at all and cannot be simply solved by stopping the development and deployment of this technology. Therefore, these systems have to be designed with enhanced safety mechanisms to minimize practices that cause the identified risks and comply with regulations.

## 5.2 EU AI Act Gap Assessment

The EU AI Act is the first comprehensive regulatory framework for artificial intelligence and serves as the current state-of-the-art. We have covered in 2.2.2 the values this

regulation aims to protect and the historical course of the concerning ethical guidelines. The EU AI Act works with a risk-based approach with its defined four levels of risks and the additional handling of general-purpose AI models with and without systematic risks. We have seen that the technology is not solely algorithm-based but is also inherently data-driven. Hence, effective regulation has to address both components, isolated and in combination. Regulating one without the other leaves an open door for exploitation.

Our gap assessment is divided into two parts; the categories that are covered by the act and suggestions to bridge open points. As defined in the methodology 3, the categories we focus on throughout our document analysis are the **technological components**, **actors**, and **risk definitions of each of the risk levels**. We primarily aim to identify if the risks we identified fall under the definitions presented by this framework. Assessing the efficiency of the obligations that this regulation poses is out of our scope.

There are a few exceptions regarding the scope of application of this regulation. The regulation does not apply to AI systems, models, and outputs that are solely used for scientific research, testing, or development (article 2 paragraph 6), and to systems and models that have not been placed on the market or deployed (article 2 paragraph 8). Testing on live instances (“real world”) does not belong to this exclusion. There is also an explicit reference to deep fakes. According to article 2 paragraph 60:

*“deep fake” means AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful.*

### 5.2.1 Categories Covered by EU AI Act

The act includes clear definitions of the different aspects concerning the regulation of AI. Among them, we found the relevant definitions for our three predefined categories.

#### Technological Components

There are several references throughout this document to the regulation of technological components, including both the **software and data**. For instance, free and open-source AI components cover both software and data. Included are models, general-purpose AI models, tools, services, or processes of an AI system (preamble 103). The security of **physical infrastructure** of AI systems is also addressed. For instance, providers need to ensure an adequate level of protection to the physical infrastructure of the model throughout its whole lifecycle (preamble 115). Nevertheless, there are no clear limits set by the act concerning what training data is admissible and what outputs can be considered as too offensive.

Further points of reference of our interest are the five points of challenges in generative AI systems discussed in 2.1.2, namely, bias, misinformation and disinformation,

data privacy, transparency, and explainability. The EU AI Act refers to each of the points, among the references are:

- **Bias:** the right to non-discrimination appears in the EU AI Act (preamble 28) as well as the sufficient representation in the training datasets that would prevent inherent biases (preamble 67, Article 10 paragraphs 2 and 3). The risk for discrimination of individuals in remote biometric identification tools, administration of justice, and democratic processes is also raised. These systems have further restrictions (preamble 54, preamble 61, preamble 70). The integration of bias detection, among the additional necessary safeguards, is also seen (preamble 70, Article 10 paragraph 5). The topic of individuals with disability and their potential algorithmic or data-driven discrimination is also mentioned (preamble 80). Human oversight and the necessity for humans-in-the-loop are also addressed and information should be provided, such that natural persons could participate in the system recommendations or decisions (Article 14 paragraph 4 point b).
- **Misinformation and Disinformation:** the EU AI Act sees the need to oblige the detection and disclosure of outputs to cope with disinformation (preamble 136). The act also emphasizes the need for embedded technical solutions to prevent the dissemination of misinformation (preamble 133). The requirement for technical robustness is key for some of the defined AI systems and should therefore be designed to avoid undesirable behaviors with errors, inconsistencies, etc. (preamble 75).
- **Privacy:** the EU AI Act emphasized the development of AI systems in accordance with privacy and data governance regulation, which includes data processing (preamble 27) throughout the whole lifecycle of the AI system (preamble 69). It is prohibited to place on the market or operate AI systems that utilize facial recognition databases harvested from the internet or CCTV (preamble 43). The access to sensitive data, e.g. health-related data, should be secured with an appropriate governance to prevent discriminatory access (preamble 68). The incorporation of anonymization (including pseudonymization) and encryption should also be included as well as algorithms that do not transmit during the training data between different parties (preamble 69, Article 10, paragraph 5 point b).
- **Transparency:** the EU AI Act sets transparency obligations based on the nature of the AI system. We analyze these transparency obligations in the following section 5.2.1, in accordance to each risk level.
- **Explainability:** the EU AI Act mentions that as part of the transparency obligations of AI systems, the necessity for traceability and explainability of these systems is also seen (preamble 27). The training datasets should include appropriate statistical properties, together with their intended use and target groups (preamble 67, Article 10 paragraph 3). We will see in 5.2.1 how explainability is achieved through the transparency obligations.

The EU AI Act and all of these mentioned references do not include technology-specific requirements, with clear matrices for measuring the success of the implementation of these obligations, or standard technologies that are considered to be sufficient.

Generally, the act addresses technological components comprehensively and provides a granular regulatory framework but only on the abstract level of general guidelines. Models that are not placed on the market yet and are in the process of development or research enjoy no regulative restrictions to encourage innovation and promote technological-scientific breakthroughs.

### Actors

The act defines explicitly the different **responsible actors and their legal responsibilities**. The regulation applies to seven stakeholders (article 2 paragraph 1), each with its defined relations to the EU. The definitions for each stakeholder are found in article 3 and include providers (article 3 paragraph 3), deployers (article 3 paragraph 4), importers (article 3 paragraph 6), distributors (article 3 paragraph 7), product manufacturers (article 3 paragraph 8), authorized representatives of providers (article 3 paragraph 5), and affected persons. In this case, social media platforms are responsible stakeholders when they provide or deploy the AI systems, but not necessarily if users use their infrastructure to disseminate AI-generated content when it was produced with external tools. Malicious parties, such as terror organizations, are mentioned only in the context of anti-terrorism mechanisms.

### Coverage of the Risk Levels

The risks are also covered by the EU AI Act. Based on our identified risks, we have three relevant risk levels and the additional aspect of general-purpose AI with systematic risks.

**Prohibited AI Practices.** The classification of prohibited AI practices that pose unacceptable risks is found in Article 5. Explicit practices are prohibited with the following relevant points to our risks:

1. It is forbidden to place on the market, deploy, or use AI system that involves subliminal techniques. In particular, these techniques imply that it is beyond the individual's consciousness or they aim to manipulate or deceive. The objective of these techniques are to lead to the distortion of individual's material or behavior. As a result, individuals would take an uninformed decision that would not be taken under other circumstances and would eventually harm significantly other individuals or groups (Article 5 Paragraph 1 Point a.).
2. It is also forbidden in this context to exploit in a targeted manner any of the vulnerabilities of individuals or groups, in particular based on their age, disability, or a specific social or economic background (Article 5 Paragraph 1 Point b).

This first point mitigates the risks of **enhancement of the in-group victim identity for terrorist causes, dehumanization, propaganda of the deed and perpetuation of violence**, and **generated terrorist propaganda based on known visual attributes** and the second point mitigates the **botnet radicalization** risk. Hence, all of our five identified unacceptable risks fall under the definition of the AI Act for prohibited AI systems.

**High-Risk AI Systems.** The classification of high-risk AI systems is found in Article 6. We found only one criterion point that is directly relevant to our risks. A high-risk AI system is considered as such, if their use is intended to directly influence an election, a referendum, or the voting behavior of individuals (Article 6 paragraph 2 point 8.b). This point partially mitigates our identified high-level risk of **fake news campaigns** and targets any malicious party that aims to influence the voting behavior of individuals.

Psychological harm is included in the potential harm that AI can cause (preamble 5). The term “psychological health” is mentioned once throughout the EU AI Act (preamble 29). This reference is done within the context of prohibited subliminal techniques and is explicitly included in the scope of significant harms:

*...whereby **significant harms**, in particular having sufficiently important adverse impacts on physical, **psychological health** or financial interests are likely to occur...*

In addition, there is an explicit reference to a higher degree of stimuli through user interfaces, e.g. virtual reality, that also largely influence and can harm significantly individuals’ behavior (preamble 29). The conclusion of this point is that these practices should be completely prohibited. Therefore, this point could mitigate our identified high-risk of **nationwide mental health problems**. However, it is not explicit enough and sets no standards for the premises that might be causing psychological harm and to what degree.

**Limited-Risk AI Systems.** The transparency obligations for providers and deployers of the limited-risk AI systems are declared in article 50. The main relevant obligations to our risks include:

1. A machine-readable marking is obligatory for AI systems with generate audio, image, video, or textual output. It is the legal responsibility of the providers (Article 50 Paragraph 2).
2. Clearly visible Human-readable disclosure is obligatory for AI outputs that include manipulated images, videos, or text. It is the legal responsibility of the deployers (Article 50 Paragraph 4). The information has to be provided in a clear and distinguishable manner with the first interaction or exposure (Article 50 Paragraph 5).

These transparency obligations outlined for generative AI models comprehensively address our concerned risks, as they explicitly encompass text-to-image and text-to-video models, the primary focus of our work. This has the secondary meaning that these obligations also cover risks of higher levels and the transparency obligations serve as the greatest lower bound for the concerning regulation of visual models. However, there are no set standards yet regarding what is a sufficient digital watermarking or concrete matrices for a sufficient disclosure. The limited risks that the transparency obligations cover include the **glorification of violence** risk, **self-radicalization within a filter bubble**, the “**othering**” within a political context, the “**Post-Truth Era**”, and the risk of **inspiration to violent acts**.

**General-Purpose AI Models and Systematic Risks** Large generative AI models with integrated capability of generating images and videos clearly fall under the definition of general-purpose AI models and are also explicitly mentioned in preamble 105. However, we have identified an important nuance. Based on the EU AI Act’s current definitions, specialized text-to-image and text-to-video models are not automatically classified as general-purpose AI due to their limited scope and lack of generality.

Considering general-purpose AI models that also include text-to-image and text-to-video scope of tasks can be considered systematic risks if they fulfill the defined criteria in article 51. In this case, each model has to be specifically evaluated and the provider has to supply the relevant parameters that play a role in determining whether the system poses systematic risks.

Certain restrictions on general-purpose AI models that do not fall under the definition of systematic risks do not hold for free and open-source models. Such models should ensure high levels of transparency, which includes published information about their weights, model architecture, and model usage (preamble 102, 104). This way, published models can be used by any digital actor.

Under general-purpose AI, we identify that based on nuances in this regulation, some models with the similar capabilities can fall under the definition of general-purpose AI systems with systematic risks that poses additional necessary restrictions and some could benefit from significantly less restrictions. We also identify the risk of duplicating models of open-source and free license software for malicious purposes that do not fall under systematic risks. This correlates with our worst-case scenario, in which the models can get out of the hands of companies that do follow the regulation and are subject to it.

### 5.2.2 Identified Gaps

Our document analysis examined pre-defined categories, showing that the EU AI Act encompasses almost all of them, namely, the technological components, the actors, and

the mitigated risks based on their levels. However, throughout our gap assessment, we have encountered several open points concerning explicit technology-specific requirements to address the primary technical challenges in generative AI models (discussed in 2.1.2) and further open points that are to be determined in the following years. We split our recommendations into technology-specific recommendations and further guidelines that concern non-technological experts.

### Technology-Specific Gaps

Throughout the analysis of the EU AI Act, the following technology-related gaps were found, considering the identified risks concerning extremist misuse and known technical challenges of generative AI systems:

**No clear standards for admissible inputs and outputs.** There are no clear set of standards for various critical points, in particular, the lack of clear matrices concerning admitted inputs and outputs presents significant challenges. There are no clear limits that define when an output becomes offensive or what training data sets shall be prohibited based on the task. A leading question, especially for non-trivial topics, is when does an output become offensive, even if intended for artistic purposes.

A trivial example that is highly relevant to our risk landscape is the use of violent training data. One possible task that is necessary for the security of social media platforms is violence recognition or other related extremism content (e.g. hate speech) and its censorship. Another possible task that stems from the same training data set is the terrorist use of generative AI and the production of the propaganda of the deed with graphic content. This task should be prohibited based on the definition of unacceptable risk. There is therefore an inherited tension between the two possible uses and no clear solutions and requirements how to isolate the training datasets for the recognition task from the generation task.

**No technological standardization and open points.** There are a few open points that are to be determined in the next two years. For example, in article 6 paragraphs 5, there will be provided guidelines with a future deadline of February 2026, or in article 56 paragraph 9 the codes of practice will be ready by May 2025. For now, there is a lack of standardization for relevant technologies concerning our risk landscape. In particular, watermarking that is a key to the transparency obligation and sufficient detection algorithms for bias detection and hate speech detection that are relevant for extremist misuse.

However, as of now, there is no technological standardization, it is not clear how the watermarking is issued, what algorithms are admissible and concerned to be secure enough for injecting watermarking or serving as extremism detection, or

whether it is necessary to have an external inspector of the issued watermarking, similar to certificate authority of digital authorities.

**Insufficient Addressing of Transparency Concerns.** We have observed a grained approach toward the transparency of AI models. Open-source AI models that do not encompass systematic risks are obliged to be transparent about their core algorithms, architecture, and used data for the pre-training while benefiting from an exemption of transparency obligations, namely no machine-readable marking or human-readable disclosure.

We do see a potential to exploit this freedom and available information that aims to promote innovation for malicious purposes instead. Therefore, we find it extremely important that companies would also be aware about our identified risks and would consider them while developing and openly publishing their software.

Furthermore, the lack of explicit requirements for human oversight places the responsibility on stakeholders' interpretations. The absence of mandated human-in-the-loop moderation weakens trust in the systems, potentially leading companies to prioritize human oversight less than necessary.

The gap analysis of the EU AI Act reveals critical technology-specific gaps, including unclear standards for admissible inputs and outputs, a lack of technological standardization, and insufficient transparency measures. These must be addressed to enhance the resilience of generative AI systems against extremist misuse, ensuring human oversight and effective technological solutions to be integrated into future iterations of the regulatory frameworks.

### Non-Technology Gaps

The gap analysis of the EU AI Act has identified several non-technological gaps that we recommend addressing in future iterations:

**No explicit reference to non-public entities as malicious stakeholders.** Our focus point in this work was terrorism and extremism. Throughout the EU AI Act, there is no explicit reference to terror organizations as a possible consumer or malicious party of AI. There is an implicit reference to these malicious parties according to article 3 paragraph 3, 4, 6, or 7. In accordance with these definitions, terror organizations can fall under the definition of *natural or legal person... or other body*.

However, we argue that the EU AI Act is a document that was primarily designed to restrict the power of public entities from taking advantage of the technology. Terrorism-related terms are only mentioned with regard to anti-terrorism acts and the focus of the act is rather on restricting the deployment of AI systems by law

enforcement in the context of anti-terrorism. Lakomy [98] identified that there is in current research focus on AI and anti-terrorism and that terrorist use of AI is rather underrepresented. Here, we identify a similar tendency with the rather implicit references to risks that stem from malicious parties, i.e. terror organizations or extremist groups.

**No media responsibility discussed.** Throughout our work, we have seen that media, both traditional media and social media platforms, serve as the infrastructure to disseminate generated content and promote its virality, also if they are not the origin platforms that outputted the content. The responsibility of these platforms not to distribute generated content without disclosure and to ensure that the disseminated information is authentic is not discussed. It is especially relevant to the transparency obligations, and as of now, it is not clear who is legally liable to enforce it. We also suggest discussing in the future the legal responsibility of media operators and their obligation to integrate deep fake identification algorithms in their platforms. We suggest a standardized approach and not a company-based policy.

**No educational programs.** We have seen in 4.5.2 that resilience is key, particularly in times of war and terrorism. Education from a young age and national programs in which individuals acquire the tools how to consume and process AI content can develop a healthy relation to the consumption of AI technologies that focus on critical thinking and can contribute to the minimization of consequent mental health problems.

We have not found any explicit reference to governmental education programs in the EU AI Act. The only reference to education is found in article 4 under “AI literacy”. According to this article, providers and deployers are mandated to ensure education and training for groups and individuals that are directly effected by AI systems. Under this definition, governments are not necessarily included. We see great importance in having a centralized approach toward education programs and suggest including this matter in future discussions.

The EU AI Act marks a significant milestone in the global regulation of artificial intelligence, establishing the first comprehensive framework aimed to balance innovation incorporating the protection of fundamental values. We have explored the act’s risk-based approach, its coverage of technological components, stakeholder definitions and responsibilities, and the categorization of risks. We have conducted a gap analysis, based on three predefined categories. The gap analysis of the EU AI Act reveals six gaps, three technology-specific and three general points. While the Act provides a foundational structure for regulating AI and largely covers our researched categories, the abstract handling of key technical values and the absence of explicit technical standards, creates an urgent necessity to establish clear technology-requirements.

### 5.3 Technical Requirements for Risk Mitigation

Our final and concluding findings address the technology-specific gaps within the legal framework of the EU AI Act, contextualized by our identified risk landscape, discussing adequate technological solutions for visual generative AI systems. For this practical contribution, it was essential to establish a solid theoretical foundation that integrates technological, social, and legal components, obtaining a holistic comprehension. The combination of these diverse perspectives represents the key contribution of our work.

To address the third research question (RQ3), we present a list of high-level technological requirements and demonstrate how they help mitigate each of the primary identified risks in 5.1. The proposed safeguards also consider the key actors, which would either prevent their misuse, in cases of direct harm, or enhance their safety and quality, in cases of indirect harm. These technology-specific solutions should be incorporated into future regulations and research to improve the security of generative AI systems and prevent extremist actors.

Our technical solutions are presented by their effectiveness in mitigating risks, starting with three primary solutions that comprehensively address the identified risks, followed by three additional technological solutions that focus on specific edge cases.

**Humans-in-the-loop Moderation.** We recommend adding explicit guidelines in future regulation regarding the integration of humans-in-the-loop after having highlighted its importance in 2.1.1. The guidelines should determine the required number of annual manual evaluations, accompanied by transparent reporting, to ensure adequate human oversight over registered and publicly available generative AI systems. Continuous integration of human input minimizes potential system biases and promotes trust and accountability through regular quality control and human-driven feedback loops for improvement.

Furthermore, the group of involved humans should be diverse, including experts from various domains such as psychology, criminology, education, and ethics, to ensure a balanced and comprehensive oversight to cover extremism-specific misuse. Each expertise contributes to the review of edge cases that are relevant to the specific domain. We have seen the importance of domain-expertise throughout the work, for instance its role in developing accurate extremism detection algorithms in 4.3.3.

**This solution inherently addresses all identified risks.** Humans-in-the-loop is a crucial technical solution for mitigating the risks of violence in systems, as it ensures essential human oversight and incorporates a diverse range of omnidomain knowledge, from social insights to technical expertise. While it cannot independently address the full spectrum of risks, it is fundamental to the effective

implementation of any technical solution, especially when tackling issues with significant societal impact. The capability to continuously integrate manual feedback enhances the transparency and explainability of these systems, which are critical for effectively managing violent content. By fostering a collaborative relationship between human judgment and automated processes, this approach helps ensure that generative AI systems remain accountable and responsive to the complexities of real-world scenarios.

**Input Validation with Keyword Detection and Semantic Analysis.** The intentional generation of extremist content can be mitigated during the processing of prompts by implementing predefined keywords and sentiment analysis methods. This proactive approach discussed in 4.3.3 can prevent problematic inputs across all types of inputs (textual, audio, and visual). We also emphasize the need for continuous development of advanced models that support cross-type inputs, addressing the lack of cross-examination techniques in detection models, as identified by Govers et al. [67]. Sentiment analysis enhances the detection of malicious and sophisticated inputs, extending the scope of harmful content identification. Establishing global libraries with labeled offensive inputs promotes human oversight, transparency, and explainability within generative AI systems.

However, the question of transparency vs. security is highly relevant for this solution, as publicly known flagging mechanisms could potentially be exploited by malicious actors who might manage to jailbreak incorporated safeguards or operate their cloned systems as we discussed in 4.4.2. The identification of user input should be performed on the server side (in case of an online API) and integrated into the generation process to minimize the risk of bypassing these security measures. To optimize the success rate of the detection algorithms, continuous integration of data and assessments by domain-experts is essential. The dictionaries have to include diverse data that represents a large scope of terms, covering all range of extremist ideologies, and reducing potential bias.

**This technological solution primarily mitigates all the risks with intended malicious intentions and direct harm.** These include:

- enhancement of the in-group victim identity for terrorist causes
- dehumanization of the out-group
- propaganda of the deed and the perpetuation of violence
- generated terrorist propaganda based on known visual attributes
- botnet radicalization
- “othering”
- fake news campaigns

The risks belonging to the radicalization category are mitigated through the use of domain-specific dictionaries and established linguistic patterns, incorporating experts input. These solutions must be designed to maximize their resilience against jailbreaking, effectively blocking in an internal layer any prompts that could trigger the generation of violent content. Additionally, input validation plays a crucial role in minimizing coordinated attempts to exploit botnet generation attacks. To address polarization risks, it is essential to remain vigilant about significant political events that involve the intentional dissemination of misinformation and disinformation.

**Automated Detection for Violent and Offensive Outputs** To prevent violent outputs, violence detection tools can serve as a second valuable safeguard following the input validation. This also intends to catch amplified violent content, as Hao et al. [72] highlighted in their research. However, it is crucial to prioritize the security and restricted access of datasets containing real-life violent content to prevent their misuse by malicious actors who might use them to train their own generative AI systems and ensure the privacy of the victims.

Additionally, there should be a strict segmentation between validation and generation tasks within the system, ensuring that harmful data does not accidentally become part of the training set. This separation would also help reduce potential bias toward violent outputs in generative AI models. Integrating this detection algorithm must follow robust security protocols, as an incomplete solution could lead to unintended and harmful outcomes.

**This key technological solution effectively addresses all identified risks.** The core issue driving our research is the generation of violent visuals through generative AI systems. Implementing a robust internal flagging mechanism as a final safeguard before delivering outputs to users can effectively mitigate the dissemination of harmful both intended and unintended content, substantially reducing the potential for harm. This solution has to be thoughtfully designed and carefully implemented, taking into the possible drawbacks and abuse potential of malicious actors.

**GAN-driven Visual Anonymization.** Anonymization algorithms effectively generalizing visual personally identifiable information (PII) of natural persons discussed in 4.3.2, particularly non-public individuals, can significantly enhance their privacy and prevent targeted attacks by malicious parties. Additionally, these algorithms can reduce the dissemination of targeted visual disinformation about non-public figures. Consequently, the training of visual generative AI using specific targeted groups of individuals could also be made more secure. Concerning our risk of psychological harm, this method could prevent edge cases of deep fakes of real individuals in atrocities or any other violent scenarios.

We recommend integrating this anonymization technique as a standard component of public visual generative AI systems, which are highly vulnerable to misuse by extremists and other malicious actors. Furthermore, we would consider establishing a database of agreed-upon public figures for whom the application of an anonymization algorithm would not be less necessary, taking into account relevant ethical considerations. This heuristic would help decrease the computational overload of this additional anonymization step.

**This technological solution primarily mitigates the following risks:**

- **Enhancement of the in-group victim identity for terrorist causes, dehumanization of the out-group, propaganda of the deed and the perpetuation of violence, botnet radicalization:** GAN-driven anonymization can serve as a powerful mechanism to disrupt radicalization processes, facilitating coordinated extremist activities. Altering facial visual attributes with each generation effectively neutralizes the intended emotional and psychological manipulation that is central to radicalization. Radicalizing content that visually reinforces in-group identity and employs dehumanization techniques on the out-group becomes less impactful when the individuals are anonymized.

This technique also combats efforts to carry out targeted recruitment or coordinated campaigns and attacks. Malicious actors lose a critical tool for creating personalized content that aims to lure individuals into extremist ideologies. This makes it harder to exploit visual propaganda to fuel the “propaganda of the deed” and perform coordinated attacks on specific groups or individuals, seen under botnet radicalization risk.

Additionally, anonymization limits the capacity of extremist networks, due to the added layer of unpredictability that disrupts the personalized designed radicalization campaigns, including manipulation of potential recruits based on visually recognizable attributes.

- **“Othering”, fake new campaigns:** the anonymization of visual attributes can majorly help combat intentionally coordinated and targeted misinformation and disinformation for political polarization. The coordination of generated visuals is completely blocked by the general anonymization taking place for each output

**Resilient Watermarking.** As discussed in 2.1.3, Nagai et al. [126] named several key requirements for effective watermarking, including fidelity, robustness, capacity, security, and efficiency. We propose adopting these principles in research for more resilient watermarking techniques. Specifically, we suggest incorporating a mechanism similar to digital signatures. This would involve an in-processing watermarking, including encrypted metadata, e.g. the generation date, non-sensitive user data (to ensure privacy) that was logged-in in the session that initiated the task to the API,

and details of the instance that generated the content (e.g. an online request or a local instance). Capturing non-sensitive user data improves safety of generative AI systems while preserving the level of privacy. The encryption would be performed using the software provider's private key and can be publicly verified through a trusted third party authorized by the EU.

To enhance accessibility, verification could be made available via tools like browser extensions, mirroring the functionality of digital certificates that confirm the legitimacy of websites. This would enhance security across the digital environment while also fostering user trust and promoting critical thinking in their consumption of digital information.

We also recommend dispersing this encrypted information across multiple, random locations within the visual content. This method helps mitigate a broad range of attack vectors, such as cropping or overwriting attacks. Also in case of compromised bits, the verification authority could still notify the user about the likelihood that the visual content is artificial.

The watermarking information should be integrated through inverted bits, which are invisible to the human eye, preserving the quality of the image. While we acknowledge the limitations of watermarking, such as the ability to erase digital traces through physical methods like printing and reproducing a digital copy via manual scanning, or the increased computational load required for secure fuzzy cryptographic integration in a randomized manner, we believe this approach can significantly enhance the security and reliability of generative content in a digital environment.

**This technological solution primarily mitigates the following risks:**

- **Propaganda for radicalization purposes** including the risks of: **propaganda of the deed and the perpetuation of violence, glorification of violence, generated terrorist propaganda based on known visual attributes and self-radicalization within a filter bubble.** Watermarking generated propaganda content designed to disseminate misinformation and disinformation can help users critically evaluate such material, allowing them to easily identify manipulation techniques. This mechanism can also greatly help in cases of self-radicalization, empowering the users to identify manipulated content.
- **“Othering”, “post-truth era”, fake news campaigns:** watermarked content contributes to the efforts of combating polarization, including “othering”, the “post-truth era”, and fake news campaigns, as it serves as a crucial tool for enhancing information transparency and integrity validation examining misinformation and disinformation. By identifying sources and flagging potentially

misleading content, watermarking empowers users to critically evaluate the credibility of the information they encounter and consume. This transparency not only aids in recognizing biased or manipulated narratives but also fosters a more informed public discourse.

Additionally, watermarking can deter the intentional spread of false information by holding creators accountable for their content via capturing their user data, handling the decentralization flow of misinformation and disinformation discussed in 4.3.2. Resilient watermarking techniques can help increase trust in media and promote healthier information consumption.

**Continuous Integration of User Feedback.** We recommend implementing a standardized system for user feedback, enabling users to report on the quality of generated content or alarm about offensive content.

This approach facilitates the traceability of harmful or offensive content, creating a pool of data that can be internally reviewed by the software provider or operator. This feature follows the principle of Humans-in-the-loop discussed in 2.1.1, by involving non-experts users, and handling their feedback with automated detection algorithms 4.3.3 that can efficiently handle vast amounts of data.

Additionally, this mechanism promotes greater trust in AI systems by incorporating and encouraging human oversight. Continuous integration of feedback and fine-tuning of generative AI systems are essential for keeping these systems accurate, responsive, and flexible, and are key to coping with misinformation and disinformation or any other present bias.

**This technological solution primarily mitigates the following risks:**

- **Dehumanization of the out-group:** victims of visual dehumanization have the ability to report offensive content. This would contribute to the optimization of visual violence detection mechanisms to correctly flag such content and prevent it from reoccurring in future iterations.
- **Fake news campaigns:** this solution could help users report fake information. The algorithms to process their queries have to be well designed to cope with multiple contradicting results that would be also reviewed manually, especially because malicious actors could abuse this feature to remove correct information.
- **Nationwide mental health problems:** this feature is extremely beneficial in reporting to software providers and operators about harmful content. This feature would play also a vital part in improving child protection in the digital space.

We found solid technical solutions that comprehensively mitigate the risks, while addressing the primary technological challenges of generative AI regarding the generation of

violent visuals.

These technological solutions can guide policymakers and researchers in prioritizing relevant mechanisms to build the resilience of generative AI systems against extremist misuse. Understanding the risk landscape is critical for correct evaluation and prioritization, which could also lead to optimized solutions, considering all risk factors. Implementing these requirements as part of the EU standards is essential to foster a safer digital environment and promote responsible use of generative AI technologies.

Our key findings are illustrated once again in the following two tables. They include the six technological solutions in combination with the five mitigated generative AI challenges and twelve identified risks. of Table 5.2, summarizes and links between the suggested technical solutions and the primary generative AI challenges discussed in 2.1.2, including bias, misinformation and disinformation, privacy, transparency, and explainability. This table provides a visual overview of the relationships discussed in the main text. In our last concluding table 5.3, we present a reverse overview that connects the identified risks outlined in 5.1 to our proposed technical solutions as part of the risk mitigation strategy.

Technological Solutions	Mitigated Generative AI Challenges
<b>Humans-in-the-loop moderation</b>	<ol style="list-style-type: none"> <li>1. bias</li> <li>2. misinformation and disinformation</li> <li>3. privacy</li> <li>4. transparency</li> <li>5. explainability</li> </ol>
<b>Input validation with keyword detection and semantic analysis</b>	<ol style="list-style-type: none"> <li>1. misinformation and disinformation</li> <li>2. transparency</li> <li>3. explainability</li> </ol>
<b>Automated detection for violent and offensive outputs</b>	<ol style="list-style-type: none"> <li>1. bias</li> <li>2. misinformation and disinformation</li> <li>3. privacy</li> <li>4. transparency</li> <li>5. explainability</li> </ol>
<b>GAN-driven visual anonymization</b>	<ol style="list-style-type: none"> <li>1. bias</li> <li>2. misinformation and disinformation</li> <li>3. privacy</li> </ol>
<b>Resilient watermarking</b>	<ol style="list-style-type: none"> <li>1. misinformation and disinformation</li> <li>2. privacy</li> <li>3. transparency</li> </ol>
<b>Continuous integration of user feedback</b>	<ol style="list-style-type: none"> <li>1. bias</li> <li>2. misinformation and disinformation</li> <li>3. transparency</li> <li>4. explainability</li> </ol>

Table 5.2: Technical Solutions to the key challenges of visual generative AI

Mitigated Risk	Technological Solutions
<b>Enhancement of the in-group victim identity for terrorist causes</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Input Validation</li> <li>3. Automated Detection for Violent Outputs</li> <li>4. GAN-driven anonymization</li> </ol>
<b>Dehumanization of the out-group</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Input Validation</li> <li>3. Automated Detection for Violent Outputs</li> <li>4. GAN-driven anonymization</li> <li>5. Continuous Integration of User Feedback</li> </ol>
<b>Propaganda of the deed and the perpetuation of violence</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Input Validation</li> <li>3. Automated Detection for Violent Outputs</li> <li>4. Resilient Watermarking</li> <li>5. GAN-driven anonymization</li> </ol>
<b>Generated terrorist propaganda based on known visual attributes</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Input Validation</li> <li>3. Automated Detection for Violent Outputs</li> <li>4. Resilient Watermarking</li> <li>5. GAN-driven anonymization</li> </ol>
<b>Glorification of violence</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Automated Detection for Violent Outputs</li> <li>3. Resilient Watermarking</li> </ol>
<b>Botnet radicalization</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Input Validation</li> <li>3. Automated Detection for Violent Outputs</li> <li>4. GAN-driven anonymization</li> <li>5. Resilient Watermarking</li> </ol>
<b>Self-radicalization within a filter bubble</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Automated Detection for Violent Outputs</li> <li>3. GAN-driven anonymization</li> <li>4. Resilient Watermarking</li> </ol>
<b>“Othering”</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Input Validation</li> <li>3. Automated Detection for Violent Outputs</li> <li>4. Resilient Watermarking</li> <li>5. GAN-driven anonymization</li> </ol>

Table 5.3: Mitigated Primary Risks and their Corresponding Technical Solutions

<b>“Post-truth era”</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Automated Detection for Violent Outputs</li> <li>3. Resilient Watermarking</li> <li>4. GAN-driven anonymization</li> </ol>
<b>Fake news campaigns</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Input Validation</li> <li>3. Automated Detection for Violent Outputs</li> <li>4. Resilient Watermarking</li> <li>5. GAN-driven anonymization</li> <li>6. Continuous Integration of User Feedback</li> </ol>
<b>Inspiration for violent acts</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Automated Detection for Violent Outputs</li> </ol>
<b>Nationwide mental health problems</b>	<ol style="list-style-type: none"> <li>1. Humans-in-the-loop Moderation</li> <li>2. Automated Detection for Violent Outputs</li> <li>3. Continuous Integration of User Feedback</li> </ol>

Table 5.4: Mitigated Primary Risks and their Corresponding Technical Solutions (cont’)

To conclude, we have defined twelve primary risks regarding the use of visual generative AI models in the violent context of war and terrorism in exacerbating radicalization, polarization, and psychological harm. We then assigned three risk levels to each identified point and discussed the relevant stakeholders, the harm each risk can cause, and the concerning harmed target groups. These findings are aligned with the most up-to-date global political occurrences and technological advancements and identify future risks to technological systems that have not been released to the public. Our defined framework highlights the risks that should be considered by companies who aim to engage in ethical development and operate systems that would minimize potential harm on such a large scale with potential global damage.

We conducted a systematic document analysis of the endorsed EU AI Act in June 2024, identifying three primary categories to evaluate the regulation’s effectiveness regarding the risks we identified. Our analysis revealed a strong correlation between these risks and the proposed mitigation strategies within the EU AI Act, with the exception of nationwide mental health issues. Notably, the absence of technology-specific standards left several key challenges only superficially addressed, lacking concrete implementation requirements. These gaps highlight the urgent need for explicit technology-related requirements to be integrated into the EU AI Act, establishing globally recognized standards.

Furthermore, we highlighted technology-focused and non-technology-focused gaps drawn from the gap analysis. These are relevant for future regulation iterations and research directions that would help mitigate our risks and minimize the potential misuse of generative AI systems. The general gaps point out the need to explicitly address terrorist actors, as malicious stakeholders, discuss media responsibility as a third-party for disseminating violent content, and promote initiatives such as national education programs to enhance

societal resilience and ensure the responsible use of generative AI technologies, fostering informed decision-making based on extensive interdisciplinary knowledge.

Lastly, we present six technological requirements that combine insights from our initial literature review in the background and the working PRISMA corpus. Notably, three solutions are remarkably effective in mitigating all identified risks, namely incorporating domain experts in content moderation, employing machine learning techniques for input validation to prevent harmful prompts, and utilizing automated detection tools for visual violence to block harmful content. Additional nuanced solutions include GAN-driven anonymization of individuals' visual attributes, resilient watermarking to alert users about generated content, and the continuous integration of user feedback. The concluding table effectively highlights our key contributions in a clear and simplified format.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# CHAPTER 6

## Conclusion

In this work, we traced the pathway through which exposure to visual violence triggers radicalization processes that serve violent extremist groups and terror organizations. The first objective of following this chain is to comprehend how generated violent content can be misused, potentially exacerbating existing negative tendencies, and to identify the associated risks it poses. The second objective is to understand whether the current regulative state-of-the-art adequately addresses these risks, in particular with sufficient technical requirements and standardization. The third and primary objective of this work is to provide explicit technical solutions and focus points for future research, that both mitigate the main challenges in generative AI systems and the risk landscape of extremist misuse and fill the gaps in the examined regulatory framework.

We developed a comprehensive understanding by applying a socio-technical lens and breaking down the different components into social and technical elements with mutual influences. This approach is highly adequate for addressing our data and research questions, allowing for a profound exploration of both technical and societal issues while illustrating their interconnections systematically.

We saw how radicalization models lay the groundwork for self-radicalization with the common elements of largely amplifying the feelings relating to the “othering” phenomenon. Furthermore, we observed how digital identities and personalized internet infrastructure facilitate radicalization in modern times and how malicious parties have taken advantage of these tendencies to structurally rebuild their communication strategy with which they lure individuals to engage in violence on their behalf. With the last layer of our socio-technical system, we portrayed the terrorist causes that encompass violence and followed their propaganda characteristics, the secondary effect of polarization that serves further actors, and the psychological consequences of exposure to graphic violence.

From this comprehensive socio-technical system, we extracted the main concerning risks by drawing the line between the discovered aspects and the known risks of visual generative AI models that we considered in the background. We described each of the risks with several components, including intended or unintended harm, risk level that correlates with the EU AI Act risk-based approach, the responsible actors, the context that has led to the risk, the general applicability of the risk with a fine-grained portrayal of the harmed target group, and the three stages of the risk ranging from the training phase to the immediate and long-term effects.

We found twelve risks, divided into the categories of radicalization, polarization, and psychological harm. We assigned the unacceptable risk level to five risks, the high-risk level to two risks, and the limited risk level to five risks. These risks encompass varied subliminal techniques integrated into propaganda, the glorification and inspiration for violent acts, the enhancement of the “othering” and “post-truth era” phenomena, and nationwide mental health problems.

We analyzed the first and most recent comprehensive regulation framework of the EU AI Act from June 2024 and examined whether it mitigates the risk landscape we defined. We found out that almost all the risks fall under the risk definitions, except the nationwide mental health problems that were only implicitly mentioned. Our identified gaps focus on both technology-specific and general points.

The lack of clear standards for technological implementations as well as admissible inputs and outputs regarding offensive content is notable. The tension regarding transparency requirements and the potential exploitation of open knowledge has also been addressed and should be reconsidered. General gaps in the EU AI Act include the lack of explicit reference to terror organizations or violent extremist groups as malicious stakeholders, the absence of established technological requirements, and the omission of media responsibility in disseminating generated content from external sources. We also strongly advocate for the inclusion of educational programs in future regulations to develop healthy consumption of AI-generated content and to provide the tools necessary for dealing with this technology, enhancing resilience in populations.

Technology-specific recommendations include the incorporation of humans in content moderation, both experts and users. Machine learning algorithms, such as predefined keywords and semantic analysis, are necessary to prevent intentional offensive prompts. We also recommend automated detection tools for visual violence that block harmful outputs, GAN-driven anonymization of individuals’ visual attributes, and the continuous integration of user feedback. Moreover, we discuss the key points for future research and development of watermarking that would lead to technological standardization, as prolonged by the EU AI Act.

The most comprehensive solutions are humans-in-the-loop and visual violence detection

---

algorithms, as they address all identified risks at once. These are therefore necessary for meeting the technological requirements of high-risk systems susceptible to extremist misuse. Input validation also plays a significant role in preventing the generation of violent content, particularly in cases involving intentional harm or attempts to bypass safeguards by malicious actors. However, it is crucial to note that no single solution is sufficient as a standalone. A combination of integrated measures is required, with each component playing a vital role in the chain of safeguards.

There are many remaining points for future work. The topic of generated sexual violence is also highly related to graphic violence and we see great importance in researching the technological capabilities to perpetuate it, examining the already implemented safety mechanisms, and finding further needed technological and regulative steps to prevent any further related harm. Delving into the effectiveness of the regulative obligations and assessing them is also an open point of our work.

Other relevant areas include the development of a secure watermarking algorithm together with our suggested approach and the jailbreak potential of the implemented safety and detection mechanisms. Further solutions for privacy-enhancing techniques in generative AI systems, in particular anonymization of individuals and the development of an agreed-upon database, are also of interest. We also see an increased need to develop matrices concerning the limits of admissible violence and training automatic detection tools that would be integrated into generative AI systems, preventing cases of intentional violence or unintentional amplified violent content.

An additional point that is out of our scope is generated violence in arts and video games, what it serves, and what restrictions in the context of these industries should be made. The last open point we wish to pursue is the ethical dilemmas of perpetuating violence through generative AI models, how it reflects our democratic values, and what future consequences it has for humankind.

We hope this research has highlighted the serious and far-reaching risks that generative AI systems pose through perpetuating violence and extremism and raised awareness of potential radicalization, polarization, and psychological harm caused by generative AI technologies. Additionally, we encourage regulators and policymakers to urgently address the identified risks and mitigate the found gaps, and establish technology-specific requirements in their future efforts toward ensuring us a more secure future, both in the physical realm and the digital space.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

## List of Figures

2.1	GAN architecture [18] . . . . .	7
2.2	Sora’s stepwise denoising process on the sequence of the patches [133] . . . . .	10
2.3	The four risk levels and the legal obligations declared by the EU AI Act . . . . .	20
3.1	The PRISMA screening steps toward a database . . . . .	28
4.1	Radicalization through Exposure to Visual Violence Pathway . . . . .	32
4.2	The interacting socio-technical components of online extremism adjusted from [21] . . . . .	33
4.3	Shin and Jitkajornwanich’s Mechanism of Feedback Loops [164] . . . . .	40
4.4	An example for the anonymization with the conditional GAN model . . . . .	42
5.1	An Overview of the Found Risks . . . . .	67
5.2	Terrorist and Extremist Risks of Misuse of Generative AI . . . . .	84
5.3	Terrorist and Extremist Risks of Misuse of Generative AI (continued) . . . . .	85

## List of Tables

2.1	The main intersecting AI ethical values [87] . . . . .	17
5.1	The Components of our Risks Framework . . . . .	69
5.2	Technical Solutions to the key challenges of visual generative AI . . . . .	100
5.3	Mitigated Primary Risks and their Corresponding Technical Solutions . . . . .	101
5.4	Mitigated Primary Risks and their Corresponding Technical Solutions (cont’) . . . . .	102



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Bibliography

- [1] Amazon scrapped 'sexist ai' tool. <https://www.bbc.com/news/technology-45809919>, 10.10.2018. Accessed on: 20.09.2024.
- [2] Sora: Aufsehenerregendes ki-video-tool von openai wird noch heuer veröffentlicht. <https://www.derstandard.at/story/3000000211684/sora-neues-zauber-tool-von-openai-wird-noch-heuer-veroeffentlicht>, 14.04.2024. Accessed on: 20.09.2024.
- [3] AGGARWAL, S., AND KUMAR, N. Attacks on blockchain. In *Advances in computers*, vol. 121. Elsevier, 2021, pp. 399–410.
- [4] AL-KHATEEB, S., AND AGARWAL, N. Examining botnet behaviors for propaganda dissemination: A case study of isil's beheading videos-based propaganda. In *2015 ieee international conference on data mining workshop (icdmw)* (2015), IEEE, pp. 51–57.
- [5] AL-SAGGAF, Y. Online radicalisation along a continuum: From when individuals express grievances to when they transition into extremism. In *Security and Privacy in Communication Networks: 14th International Conference, SecureComm 2018, Singapore, Singapore, August 8-10, 2018, Proceedings, Part II* (2018), Springer, pp. 429–440.
- [6] AL-ZZAWI, A. F. K. Impact of terrorism act on child psychology and post-traumatic stress disorder. *EXECUTIVE EDITOR* 10, 1 (2019), 298.
- [7] ALDERA, S., EMAM, A., AL-QURISHI, M., ALRUBAIAN, M., AND ALOTHAIM, A. Online extremism detection in textual content: a systematic literature review. *IEEE Access* 9 (2021), 42384–42396.
- [8] ALY, A., MACDONALD, S., JARVIS, L., AND CHEN, T. M. Introduction to the special issue: Terrorist online propaganda and radicalization, 2017.
- [9] ANANTRASIRICHAI, N., AND BULL, D. Artificial intelligence in the creative industries: a review. *Artificial intelligence review* 55, 1 (2022), 589–656.

- [10] ANDERSEN, J. C., AND SANDBERG, S. Islamic state propaganda: Between social movement framing and subcultural provocation. *Terrorism and Political Violence* 32, 7 (2020), 1506–1526.
- [11] ASONGU, S. A., ORIM, S.-M. I., AND NTING, R. T. Terrorism and social media: global evidence. *Journal of Global Information Technology Management* 22, 3 (2019), 208–228.
- [12] BAISLEY, E. Genocide and constructions of hutu and tutsi in radio propaganda. *Race & Class* 55, 3 (2014), 38–59.
- [13] BALDASSARRI, D., AND GELMAN, A. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology* 114, 2 (2008), 408–446.
- [14] BANNOUR, N., GHANNAY, S., NÉVÉOL, A., AND LIGOZAT, A.-L. Evaluating the carbon footprint of nlp methods: a survey and analysis of existing tools. In *Proceedings of the second workshop on simple and efficient natural language processing* (2021), pp. 11–21.
- [15] BAUGUT, P., AND NEUMANN, K. Describing perceptions of media influence among radicalized individuals: The case of jihadists and non-violent islamists. *Political Communication* 37, 1 (2020), 65–87.
- [16] BAUGUT, P., AND NEUMANN, K. Online news media and propaganda influence on radicalized individuals: Findings from interviews with islamist prisoners and former islamists. *New Media & Society* 22, 8 (2020), 1437–1461.
- [17] BAUMANN, F., LORENZ-SPREEN, P., SOKOLOV, I. M., AND STARNINI, M. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters* 124, 4 (2020), 048301.
- [18] BENGESI, S., EL-SAYED, H., SARKER, M. K., HOUKPATI, Y., IRUNGU, J., AND OLADUNNI, T. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers. *IEEE Access* (2024).
- [19] BIRD, C., UNGLESS, E., AND KASIRZADEH, A. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2023), pp. 396–410.
- [20] BORUM, R. The etiology of radicalization. *The handbook of the criminology of terrorism* (2016), 15–32.
- [21] BOSTROM, R. P., AND HEINEN, J. S. Mis problems and failures: A socio-technical perspective. part i: The causes. *MIS quarterly* (1977), 17–32.
- [22] BOWEN, G. A. Document analysis as a qualitative research method. *Qualitative research journal* 9, 2 (2009), 27–40.

- [23] BOYD, R. W., AND SWANSON, W. S. The evolution of virtual violence: how mobile screens provide windows to real violence. *Pediatrics* 138, 2 (2016).
- [24] BRAJAWIDAGDA, U., REDDICK, C. G., AND CHATFIELD, A. T. Social media and urban resilience: A case study of the 2016 jakarta terror attack. In *Proceedings of the 17th International digital government research conference on digital government research* (2016), pp. 445–454.
- [25] BRIGGS, R. O., NUNAMAKER, J. F., AND SPRAGUE, R. H. Social aspects of sociotechnical systems. *Journal of Management Information Systems* 27, 1 (2010), 13–16.
- [26] BROWN, O., SMITH, L. G., DAVIDSON, B. I., RACEK, D., AND JOINSON, A. Online signals of extremist mobilization. *Personality and Social Psychology Bulletin* (2024), 01461672241266866.
- [27] BRYANS, S. *Handbook on the management of violent extremist prisoners and the prevention of radicalization to violence in prisons*. United Nations Office on Drugs and Crime, 2016.
- [28] CACIOPPO, J. T., REIS, H. T., AND ZAUTRA, A. J. Social resilience: the value of social fitness with an application to the military. *American Psychologist* 66, 1 (2011), 43.
- [29] CALLANDER, S., AND CARBAJAL, J. C. Cause and effect in political polarization: A dynamic analysis. *Journal of Political Economy* 130, 4 (2022), 825–880.
- [30] CAO, Y., LI, S., LIU, Y., YAN, Z., DAI, Y., YU, P. S., AND SUN, L. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226* (2023).
- [31] CHARLTON, A., AND LEE, M. Us decides to rejoin unesco and pay back dues, to counter chinese influence. <https://apnews.com/article/unesco-us-rejoin-palestine-china-5b7849bd2cae966e4e9837380c0c094f>, 12.06.2023. Accessed on: 20.09.2024.
- [32] CHEN, X. Social media in the post-truth and political manipulation era: Let’s re-debate michael gorman vs. web 2.0. *Internet Reference Services Quarterly* 25, 1-2 (2021), 1–7.
- [33] CHUA, Y. T. “we want you!” applying social network analysis to online extremist communities. *Terrorism and Political Violence* (2024), 1–15.
- [34] COHNITZ, D., AND RAUB, W. Workshop on polarisation and radicalisation in social systems, theoretical models and empirical research. *Utrecht University* (2018).
- [35] CONWAY, M., AND DILLON, J. Future trends: Live-streaming terrorist attacks. *VOX-Pol*. Accessed October 8 (2019).

- [36] CORN-REVERE, R. I will defend to the death your right to say it: But how. *GPSolo* 35 (2018), 68.
- [37] CORRÊA, N. K., GALVÃO, C., SANTOS, J. W., DEL PINO, C., PINTO, E. P., BARBOSA, C., MASSMANN, D., MAMBRINI, R., GALVÃO, L., TEREM, E., ET AL. Worldwide ai ethics: A review of 200 guidelines and recommendations for ai governance. *Patterns* 4, 10 (2023).
- [38] COSTANZA, W. A. *An interdisciplinary framework to assess the radicalization of youth towards violent extremism across cultures*. Georgetown University, 2012.
- [39] CRESSWELL, J. W. Qualitative, quantitative. and mixed methods approaches. *Research Design*, 10, 08941939.2012 (2003), 723954.
- [40] DALTON, R. J. Generational change in elite political beliefs: The growth of ideological polarization. *The Journal of Politics* 49, 4 (1987), 976–997.
- [41] DE ZAYAS, D., AND MATUSITZ, J. Understanding the dissemination of isis beheading videos through the diffusion of innovations (doi) theory. *Journal of policing, intelligence and counter terrorism* 16, 3 (2021), 205–222.
- [42] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [43] DIMITROFF, K. Mark zuckerberg, joe manchin, and isis: What facebook’s international terrorism lawsuits can teach us about the future of section 230 reform. *Tex. L. Rev.* 100 (2021), 153.
- [44] DING, M., RABBANI, T., AN, B., AGRAWAL, A., XU, Y., DENG, C., ZHU, S., MOHAMED, A., WEN, Y., GOLDSTEIN, T., ET AL. Waves: Benchmarking the robustness of image watermarks. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.
- [45] DOOSJE, B., MOGHADDAM, F. M., KRUGLANSKI, A. W., DE WOLF, A., MANN, L., AND FEDDES, A. R. Terrorism, radicalization and de-radicalization. *Current Opinion in Psychology* 11 (2016), 79–84.
- [46] DOUAI, A. 23 technology and terrorism: Media symbiosis and the “dark side” of the web. *Communication and technology* 5 (2015), 439.
- [47] DURKHEIM, E. *Elementary forms of religious life: The totémic system in Australia*. Preface by Jean-Paul Willaime. Puf, 2014.
- [48] DURKHEIM, E. The division of labor in society. In *Social stratification*. Routledge, 2018, pp. 217–222.

- [49] ERIKSSON KRUTRÖK, M., AND LINDGREN, S. Social media amplification loops and false alarms: Towards a sociotechnical understanding of misinformation during emergencies. *The Communication Review* 25, 2 (2022), 81–95.
- [50] ESMAILZADEH, Y., AND MOTAGHI, E. International terrorism and social threats of artificial intelligence. *Journal of Globalization Studies* 15 (05 2024), 168–179.
- [51] EUROPEAN COMMISSION. Ethics guidelines for trustworthy ai. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>, April 2019. Accessed on: 16.10.2024.
- [52] EUROPEAN PARLIAMENT AND COUNCIL OF THE EUROPEAN UNION. Artificial intelligence act regulation 2024/1689 of the european parliament and of the council. [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689), 13.06.2024. Accessed on: 16.10.2024.
- [53] FERNANDEZ, M., ASIF, M., AND ALANI, H. Understanding the roots of radicalisation on twitter. In *Proceedings of the 10th ACM conference on web science* (2018), pp. 1–10.
- [54] FERRARA, E. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science* (2024), 1–21.
- [55] FIORINA, M. P., AND ABRAMS, S. J. Political polarization in the american public. *Annu. Rev. Polit. Sci.* 11 (2008), 563–588.
- [56] FLOREA, M. Media violence and the cathartic effect. *Procedia-social and behavioral sciences* 92 (2013), 349–353.
- [57] FOR ECONOMIC CO-OPERATION, O., AND DEVELOPMENT. Oecd ai principles. <https://oecd.ai/en/ai-principles>, May 2019. Accessed on: 16.10.2024.
- [58] FORUM, W. E. The global risks report 2024. [https://www3.weforum.org/docs/WEF\\_The\\_Global\\_Risks\\_Report\\_2024.pdf](https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf), January 2024. Accessed on: 16.10.2024.
- [59] FRIIS, S. M. ‘beyond anything we have ever seen’: beheading videos and the visibility of violence in the war against isis. *International Affairs* 91, 4 (2015), 725–746.
- [60] G20. G20 ai principles for responsible stewardship of trustworthy ai. <https://www.g20.org/en/g20-summit/osaka-2019/g20-ai-principles/>, June 2019. Accessed on: 16.10.2024.
- [61] GAIKWAD, M., AHIRRAO, S., PHANSALKAR, S., AND KOTECHA, K. Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *Ieee Access* 9 (2021), 48364–48404.

- [62] GARCIA, D., AND RIMÉ, B. Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological science* 30, 4 (2019), 617–628.
- [63] GOKTAS, P. Ethics, transparency, and explainability in generative ai decision-making systems: a comprehensive bibliometric study. *Journal of Decision Systems* (2024), 1–29.
- [64] GOLDSTEIN, J. S., AND PEVEHOUSE, J. C. International relations: Eight edition, 2008.
- [65] GOLDSTEIN, R. B., SMITH, S. M., CHOU, S. P., SAHA, T. D., JUNG, J., ZHANG, H., PICKERING, R. P., RUAN, W. J., HUANG, B., AND GRANT, B. F. The epidemiology of dsm-5 posttraumatic stress disorder in the united states: results from the national epidemiologic survey on alcohol and related conditions-iii. *Social psychiatry and psychiatric epidemiology* 51 (2016), 1137–1148.
- [66] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144.
- [67] GOVERS, J., FELDMAN, P., DANT, A., AND PATROS, P. Down the rabbit hole: Detecting online extremism, radicalisation, and politicised hate speech. *ACM Computing Surveys* 55, 14s (2023), 1–35.
- [68] GRIZZARD, M., HUANG, J., WEISS, J. K., NOVOTNY, E. R., FITZGERALD, K. S., AHN, C., NGOH, Z., PLANTE, A., AND CHU, H. Graphic violence as moral motivator: The effects of graphically violent content in news. In *Media, Terrorism and Society*. Routledge, 2020, pp. 29–49.
- [69] GUNTON, K. The impact of the internet and social media platforms on radicalisation to terrorism and violent extremism. In *Privacy, Security And Forensics in The Internet of Things (IoT)*. Springer, 2022, pp. 167–177.
- [70] HAMED, A. A., ZACHARA-SZYMANSKA, M., AND WU, X. Safeguarding authenticity for mitigating the harms of generative ai: Issues, research agenda, and policies for detection, fact-checking, and ethical ai. *IScience* (2024).
- [71] HAMIEL, D., WOLMER, L., SPIRMAN, S., AND LAOR, N. Comprehensive child-oriented preventive resilience program in israel based on lessons learned from communities exposed to war, terrorism and disaster. In *Child & Youth Care Forum* (2013), vol. 42, Springer, pp. 261–274.
- [72] HAO, S., SHELBY, R., LIU, Y., SRINIVASAN, H., BHUTANI, M., AYAN, B. K., PODDAR, S., AND LASZLO, S. Harm amplification in text-to-image models. *arXiv preprint arXiv:2402.01787* (2024).

- [73] HERRMAN, H., STEWART, D. E., DIAZ-GRANADOS, N., BERGER, E. L., JACKSON, B., AND YUEN, T. What is resilience? *The Canadian Journal of Psychiatry* 56, 5 (2011), 258–265.
- [74] HOLMAN, E. A., GARFIN, D. R., AND SILVER, R. C. It matters what you see: Graphic media images of war and terror may amplify distress. *Proceedings of the National Academy of Sciences* 121, 29 (2024), e2318465121.
- [75] HOLT, T. J., FREILICH, J. D., AND CHERMAK, S. M. Examining the online expression of ideology among far-right extremist forum users. *Terrorism and Political Violence* 34, 2 (2022), 364–384.
- [76] HORGAN, J. From profiles to pathways and roots to routes: Perspectives from psychology on radicalization into terrorism. *The ANNALS of the American Academy of Political and Social Science* 618, 1 (2008), 80–94.
- [77] HOSSEINMARDI, H., GHASEMIAN, A., CLAUSET, A., MOBIUS, M., ROTHSCHILD, D. M., AND WATTS, D. J. Examining the consumption of radical content on youtube. *Proceedings of the National Academy of Sciences* 118, 32 (2021), e2101967118.
- [78] HUESMANN, L. R., AND TAYLOR, L. D. The role of media violence in violent behavior. *Annu. Rev. Public Health* 27 (2006), 393–415.
- [79] HUKKELÅS, H., MESTER, R., AND LINDSETH, F. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing* (2019), Springer, pp. 565–578.
- [80] HUNTER, L. Y., BIGLAISER, G., MCGAUVRAN, R. J., AND COLLINS, L. The effects of social media on domestic terrorism. *Behavioral Sciences of Terrorism and Political Aggression* (2022), 1–25.
- [81] HUNTER, T. Ai porn is easy to make now. for women, that’s a nightmare. *The Washington Post* (2023), NA–NA.
- [82] HWANG, J., AND OH, S. A brief survey of watermarks in generative ai. In *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)* (2023), IEEE, pp. 1157–1160.
- [83] IBRAHIM, Y. Livestreaming the ‘wretched of the earth’: The christchurch massacre and the ‘death-bound subject’. *Ethnicities* 20, 5 (2020), 803–822.
- [84] IEEE GLOBAL INITIATIVE ON ETHICS OF AUTONOMOUS AND INTELLIGENT SYSTEMS. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. [https://standards.ieee.org/wp-content/uploads/import/documents/other/ead\\_v2.pdf](https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf), December 2016. Version 1 - For Public Discussion.

- [85] IVIANSKY, Z. Individual terror: Concept and typology. *Journal of Contemporary History* 12, 1 (1977), 43–63.
- [86] JAMIESON, K. H., AND CAPPELLA, J. N. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.
- [87] JOBIN, A., IENCA, M., AND VAYENA, E. The global landscape of ai ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.
- [88] JONG, W., AND DÜCKERS, M. L. Self-correcting mechanisms and echo-effects in social media: An analysis of the “gunman in the newsroom” crisis. *Computers in Human Behavior* 59 (2016), 334–341.
- [89] JORDAN, T. Does online anonymity undermine the sense of personal responsibility? *Media, Culture & Society* 41, 4 (2019), 572–577.
- [90] JURAFSKY, D., AND MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. 2024. Online manuscript released August 20, 2024.
- [91] KAPLAN, A. The psychodynamics of terrorism. *Studies in Conflict & Terrorism* 1, 3-4 (1978), 237–254.
- [92] KNIBBELER, D., AND ZADEH, S. International: The interplay between the ai act and the gdpr - ai series part 1. <https://www.dataguidance.com/opinion/international-interplay-between-ai-act-and-gdpr-ai>, November 2023. Accessed on: 20.09.2024.
- [93] KRUGLANSKI, A. W., BÉLANGER, J. J., AND GUNARATNA, R. Other Theories of Radicalization. In *The Three Pillars of Radicalization: Needs, Narratives, and Networks*. Oxford University Press, 06 2019.
- [94] KRUGLANSKI, A. W., GELFAND, M. J., BÉLANGER, J. J., SHEVELAND, A., HETIARACHCHI, M., AND GUNARATNA, R. The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology* 35 (2014), 69–93.
- [95] KRUGLOVA, A. For god, for tsar and for the nation: authenticity in the russian imperial movement’s propaganda. *Studies in Conflict & Terrorism* 47, 6 (2024), 645–667.
- [96] KÜEY, L. Role of media/social media in aftermath of violent acts/terror attacks. In *Risk Management of Terrorism Induced Stress*. IOS Press, 2020, pp. 152–162.
- [97] KÜHN, S., KUGLER, D. T., SCHMALEN, K., WEICHENBERGER, M., WITT, C., AND GALLINAT, J. Does playing violent video games cause aggression? a longitudinal intervention study. *Molecular psychiatry* 24, 8 (2019), 1220–1234.

- [98] LAKOMY, M. Artificial intelligence as a terrorism enabler? understanding the potential impact of chatbots and image generators on online terrorist activities. *Studies in Conflict & Terrorism* (2023), 1–21.
- [99] LAQUEUR, W. *A history of terrorism*. Routledge, 2017.
- [100] LEISTEDT, S. J. On the radicalization process. *Journal of forensic sciences* 61, 6 (2016), 1588–1591.
- [101] LEVAOT, Y. *Social Media Use and its Relations with Posttraumatic Stress, Post-traumatic Growth and Wellbeing Following Exposure to Traumatic Event*. PhD thesis, University of Haifa (Israel), 2021.
- [102] LEVI-BELZ, Y., GROWEISS, Y., BLANK, C., AND NERIA, Y. Ptsd, depression, and anxiety after the october 7, 2023 attack in israel: a nationwide prospective study. *EClinicalMedicine* 68 (2024).
- [103] LEVIN, Y., BAR-OR, R. L., FORER, R., VASERMAN, M., KOR, A., AND LEVRAN, S. The association between type of trauma, level of exposure and addiction. *Addictive behaviors* 118 (2021), 106889.
- [104] LIBERATI, A., ALTMAN, D. G., TETZLAFF, J., MULROW, C., GÖTZSCHE, P. C., IOANNIDIS, J. P., CLARKE, M., DEVEREAUX, P. J., KLEIJNEN, J., AND MOHER, D. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of internal medicine* 151, 4 (2009), W–65.
- [105] LIU, Y., ZHANG, K., LI, Y., YAN, Z., GAO, C., CHEN, R., YUAN, Z., HUANG, Y., SUN, H., GAO, J., ET AL. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177* (2024).
- [106] LOBINGER, K., AND BRANTNER, C. Niemand muss diese videos zeigen. der medienethische diskurs über die visuelle berichterstattung zum terroranschlag 2020 in wien.
- [107] LORENZ, P., PERSET, K., AND BERRYHILL, J. Initial policy considerations for generative artificial intelligence.
- [108] LU, Y., AND EBRAHIMI, T. Assessment framework for deepfake detection in real-world situations. *EURASIP Journal on Image and Video Processing* 2024, 1 (2024), 6.
- [109] MACKLIN, G. The christchurch attacks: Livestream terror in the viral video age. *CtC Sentinel* 12, 6 (2019), 18–29.
- [110] MAESELEE, P. A., VERLEYE, G., STEVENS, I., AND SPECKHARD, A. Psychosocial resilience in the face of a mediated terrorist threat. *Media, War & Conflict* 1, 1 (2008), 50–69.

- [111] MAKHORTYKH, M., ZUCKER, E. M., SIMON, D. J., BULTMANN, D., AND ULLOA, R. Shall androids dream of genocides? how generative ai can change the future of memorialization of mass atrocities. *Discover Artificial Intelligence* 3, 1 (2023), 28.
- [112] MAKREHCHI, M. The correlation between language shift and social conflicts in polarized social media. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (2014), vol. 2, IEEE, pp. 166–171.
- [113] MASON, L. *Uncivil agreement: How politics became our identity*. University of Chicago Press, 2018.
- [114] MCCAULEY, C., AND MOSKALENKO, S. Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and political violence* 20, 3 (2008), 415–433.
- [115] MCCAULEY, C., AND MOSKALENKO, S. Individual and group mechanisms of radicalization. *Protecting the homeland from international and domestic terrorism threats: Current multi-disciplinary perspectives on root causes, the role of ideology, and programs for counter-radicalisation and disengagement* 82 (2010).
- [116] MCCOY, J., RAHMAN, T., AND SOMER, M. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist* 62, 1 (2018), 16–42.
- [117] MCELREATH, D. H., DOSS, D. A., MCELREATH, L., LINDSLEY, A., LUSK, G., SKINNER, J., AND WELLMAN, A. The communicating and marketing of radicalism: A case study of isis and cyber recruitment. *International Journal of Cyber Warfare and Terrorism (IJCWT)* 8, 3 (2018), 26–45.
- [118] MCKEOWN, S., HAJI, R., BRYANT, J., DELA PAZ, E., AND FLOTHMANN, C. (de) humanization of muslim immigrants: Newspaper discourse and public responses during the uk 2015 general election. *The Psychology of Political Behavior in a Time of Change* (2021), 575–595.
- [119] MCNEIL-WILLSON, R., GERRAND, V., SCRINZI, F., AND TRIANDAFYLLIDOU, A. Polarisation, violent extremism and resilience in europe today: An analytical framework. Tech. rep., BRaVE Project, 2019.
- [120] MELTON, G. B., AND SIANKO, N. How can government protect mental health amid a disaster? *American Journal of Orthopsychiatry* 80, 4 (2010), 536.
- [121] MILLER, B., AND RECORD, I. Justified belief in a digital age: On the epistemic implications of secret internet technologies. *Episteme* 10, 2 (2013), 117–134.
- [122] MILTON, D. Truth and lies in the caliphate: The use of deception in islamic state propaganda. *Media, War & Conflict* 15, 2 (2022), 221–237.

- [123] MRUG, S., MADAN, A., COOK, E. W., AND WRIGHT, R. A. Emotional and physiological desensitization to real-life and movie violence. *Journal of youth and adolescence* 44 (2015), 1092–1108.
- [124] MUGHAL, R., DEMARINIS, V., NORDENDAHL, M., LONE, H., PHILLIPS, V., AND BOYD-MACMILLAN, E. Public mental health approaches to online radicalisation: an empty systematic review. *International journal of environmental research and public health* 20, 16 (2023), 6586.
- [125] MUMTAZ, N., EJAZ, N., HABIB, S., MOHSIN, S. M., TIWARI, P., BAND, S. S., AND KUMAR, N. An overview of violence detection techniques: current challenges and future directions. *Artificial intelligence review* 56, 5 (2023), 4641–4666.
- [126] NAGAI, Y., UCHIDA, Y., SAKAZAWA, S., AND SATOH, S. Digital watermarking for deep neural networks. *International Journal of Multimedia Information Retrieval* 7 (2018), 3–16.
- [127] NAPOLEONI, L. The evolution of terrorist financing since 9/11: How the new generation of jihadists fund themselves. In *Terronomics*. Routledge, 2016, pp. 13–25.
- [128] NERIA, Y., WICKRAMARATNE, P., OLFSON, M., GAMEROFF, M. J., PILOWSKY, D. J., LANTIGUA, R., SHEA, S., AND WEISSMAN, M. M. Mental and physical health consequences of the september 11, 2001 (9/11) attacks in primary care: a longitudinal study. *Journal of traumatic stress* 26, 1 (2013), 45–55.
- [129] NEUMANN, P. R. The trouble with radicalization. *International affairs* 89, 4 (2013), 873–893.
- [130] NGUYEN, C. T. Echo chambers and epistemic bubbles. *Episteme* 17, 2 (2020), 141–161.
- [131] OLSON, I. Too ‘extreme’: gonzo, snuff, and governmentality. *Porn Studies* 3, 4 (2016), 398–410.
- [132] O’NEIL, C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.
- [133] OPENAI. Video generation models as world simulators: Sora’s technical report. <https://openai.com/index/video-generation-models-as-world-simulators/>, 2024. Accessed on: 16.10.2024.
- [134] O’SHAUGHNESSY, N. J. *Politics and propaganda: Weapons of mass seduction*. Manchester University Press, 2004.
- [135] OSLER, L., AND ZAHAVI, D. Sociality and embodiment: Online communication during and after covid-19. *Foundations of Science* 28, 4 (2023), 1125–1142.

- [136] PARISER, E. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.
- [137] PASHENTSEV, E. N., AND BAZARKINA, D. Y. Isis propaganda on the internet, and effective counteraction. *Journal of Political Marketing* 20, 1 (2021), 17–33.
- [138] PEROV, V., AND PEROVA, N. Ai hallucinations: Is “artificial evil” possible? In *2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT)* (2024), IEEE, pp. 114–117.
- [139] PICART, C. J. S. “jihad cool/jihad chic”: The roles of the internet and imagined relations in the self-radicalization of colleen larose (jihad jane). *Societies* 5, 2 (2015), 354–383.
- [140] POURNARAS, E. Decentralization in digital societies—a design paradox. *arXiv preprint arXiv:2001.01511* (2020).
- [141] PRIMORATZ, I. What is terrorism? *Journal of applied philosophy* 7, 2 (1990), 129–138.
- [142] PROCTOR, D. Cybernetics and digital whiteness: Exposure to radicalization through feedback loops. In *2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW)* (2021), IEEE, pp. 1–5.
- [143] PUCZYŃSKA, J., PODHAJSKI, M., WOJTASIK, K., AND TOMASZ, P. M. Large language models in jihadist terrorism and crimes. *Terroryzm. Studia, analizy, prewencja* (2024), 351–370.
- [144] PULIDO, M. L. The ripple effect: Lessons learned about secondary traumatic stress among clinicians responding to the september 11th terrorist attacks. *Clinical Social Work Journal* 40 (2012), 307–315.
- [145] RABINOVICH, M. Psychodynamic emotional regulation in view of wolpe’s desensitization model. *The American Journal of Psychology* 129, 1 (2016), 65–79.
- [146] RAMZAN, M., ABID, A., KHAN, H. U., AWAN, S. M., ISMAIL, A., AHMED, M., ILYAS, M., AND MAHMOOD, A. A review on state-of-the-art violence detection techniques. *IEEE Access* 7 (2019), 107560–107575.
- [147] RANAKOTI, P., YADAV, S., APURVA, A., TOMER, S., AND ROY, N. R. Deep web & online anonymity. In *2017 International conference on computing and communication technologies for smart nation (IC3TSN)* (2017), IEEE, pp. 215–219.
- [148] RARM, L. Terror: live. *Continuum* 37, 3 (2023), 422–432.
- [149] RAVI, M., NEGI, A., AND CHITNIS, S. A comparative review of expert systems, recommender systems, and explainable ai. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)* (2022), IEEE, pp. 1–8.

- [150] RISIUS, M., BLASIAK, K. M., WIBISONO, S., JABRI-MARKWELL, R., AND LOUIS, W. Dynamic matrix of extremisms and terrorism (dmet): a continuum approach towards identifying different degrees of extremisms. *arXiv preprint arXiv:2312.00337* (2023).
- [151] RISIUS, M., BLASIAK, K. M., WIBISONO, S., AND LOUIS, W. R. The digital augmentation of extremism: Reviewing and guiding online extremism research from a sociotechnical perspective. *Information Systems Journal* 34, 3 (2024), 931–963.
- [152] ROBERTS-INGLESON, E. M., AND MCCANN, W. S. The link between misinformation and radicalisation. *Perspectives on Terrorism* 17, 1 (2023), 36–49.
- [153] ROBINSON, N., AND WHITTAKER, J. Playing for hate? extremism, terrorism, and videogames. *Studies in Conflict & Terrorism* (2020), 1–36.
- [154] ROSE, S. The isis propaganda war: a hi-tech media jihad. *The Guardian* 7 (2014).
- [155] ROUSSEAU, C., AGGARWAL, N. K., AND KIRMAYER, L. J. Radicalization to violence: A view from cultural psychiatry, 2021.
- [156] RYCHWALSKA, A., AND ROSZCZYŃSKA-KURASIŃSKA, M. Polarization on social media: when group dynamics leads to societal divides.(2018). DOI: <https://doi.org/10.24251/hicss> (2018).
- [157] SAAD, A., AND ALHUMAID, A. How terrorist groups use social networking media to attract and recruit new members. In *2018 International Conference on Innovations in Information Technology (IIT)* (2018), IEEE, pp. 129–134.
- [158] SABOUNI, S., CULLEN, A., AND ARMITAGE, L. A preliminary radicalisation framework based on social engineering techniques. In *2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)* (2017), IEEE, pp. 1–5.
- [159] SAGEMAN, M. The next generation of terror. *Foreign policy*, 165 (2008), 37.
- [160] SAGEMAN, M. *Leaderless jihad: Terror networks in the twenty-first century*. University of Pennsylvania Press, 2011.
- [161] SAN BIAGIO, M., SIMONCINI, S., LA MATTINA, E., AND MORREALE, V. Marple: A framework for social media threat intelligence. In *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)* (2024), IEEE, pp. 1–6.
- [162] SANTOS, F. P., LELKES, Y., AND LEVIN, S. A. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences* 118, 50 (2021), e2102141118.

- [163] SCHULTE-SASSE, L. Leni riefenstahl's feature films and the question of a fascist aesthetic. *Cultural Critique*, 18 (1991), 123–148.
- [164] SHIN, D., AND JITKAJORNWANICH, K. How algorithms promote self-radicalization: Audit of tiktok's algorithm using a reverse engineering method. *Social Science Computer Review* 42, 4 (2024), 1020–1040.
- [165] SHORTLAND, N., NADER, E., IMPERILLO, N., ROSS, K., AND DMELLO, J. The interaction of extremist propaganda and anger as predictors of violent responses. *Journal of interpersonal violence* 36, 3-4 (2021), NP1391–1411NP.
- [166] SHREE, T., AND GUPTA, S. Role of social media in online radicalization: Literature review and research agenda. *Asia pacific journal of information systems* 29, 2 (2019), 268–282.
- [167] SHWEDER, R. A., HAIDT, J., AND HORTON, R. and craig joseph. *Handbook of Emotions* (2008), 409.
- [168] SIDDIKA, A. Media violence: A study. *Elementary Education Online* 19, 3 (2022), 4851–4851.
- [169] SINGER, U., POLYAK, A., HAYES, T., YIN, X., AN, J., ZHANG, S., HU, Q., YANG, H., ASHUAL, O., GAFNI, O., ET AL. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
- [170] SOBOL, I. Glorification of Terrorist Violence at the European Court of Human Rights. *Human Rights Law Review* 24, 3 (06 2024), ngae017.
- [171] STEVENSON, J., EDWARDS, M., AND RASHID, A. Analysing the activities of far-right extremists on the parler social network. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining* (2023), pp. 392–399.
- [172] SUGARA, R., ULFA, M., AGUSTIN, F. E. M., SUBCHI, I., FARIDA, A. R., AND TOYIBAH, D. The use of electronic communication technology by terrorist movement in indonesia. In *2022 International Conference on Science and Technology (ICOSTECH)* (2022), IEEE, pp. 1–6.
- [173] SULLIVAN, A., AND MONTASARI, R. The use of the internet and the internet of things in modern terrorism and violent extremism. In *Privacy, security and forensics in the internet of things (IoT)*. Springer, 2022, pp. 151–165.
- [174] SULTANI, W., CHEN, C., AND SHAH, M. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 6479–6488.

- [175] TAKALE, D. G., MAHALLE, P. N., AND SULE, B. Exploring watermarking techniques in generative ai: A brief overview. *Journal of Image Processing and Artificial Intelligence* 10, 3 (2024), 1–5.
- [176] TEAM, T. A. E. 15 leading cloud providers for gpu-powered llm fine-tuning and training. <https://towardsai.net/p/machine-learning/15-leading-cloud-providers-for-gpu-powered-llm-fine-tuning-and-training>, 23.12.2023. Accessed on: 20.09.2024.
- [177] THOMPSON, G. Parallels in propaganda? a comparative historical analysis of islamic state and the nazi party. *Journal of Public relations research* 29, 1 (2017), 51–66.
- [178] TÖRNBERG, P., AND TÖRNBERG, A. Inside a white power echo chamber: Why fringe digital spaces are polarizing politics. *New Media & Society* 26, 8 (2024), 4511–4533.
- [179] UNESCO. Recommendation on the ethics of artificial intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000380455>, November 2021. Accessed on: 16.10.2024.
- [180] VASWANI, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
- [181] VENKATESH, V., PODOSHEN, J. S., WALLIN, J., RABAH, J., AND GLASS, D. Promoting extreme violence: visual and narrative analysis of select ultraviolent terror propaganda videos produced by the islamic state of iraq and syria (isis) in 2015 and 2016. *Terrorism and political violence* 32, 8 (2020), 1753–1775.
- [182] VOJÍŘ, S., KUČERA, J., ET AL. Towards re-decentralized future of the web: Privacy, security and technology development. *Acta Informatica Pragensia* 10, 3 (2021), 349–369.
- [183] VON BEHR, I. Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism.
- [184] WANG, Y.-C., XUE, J., WEI, C., AND KUO, C.-C. J. An overview on generative ai at scale with edge-cloud computing. *IEEE Open Journal of the Communications Society* (2023).
- [185] WEBBER, D., AND KRUGLANSKI, A. W. Psychological factors in radicalization: A “3 n” approach. *The handbook of the criminology of terrorism* (2016), 33–46.
- [186] WIBAWA, D. Media construction and radicalism. In *International Conference on Media and Communication Studies (ICOMACS 2018)* (2018), Atlantis Press, pp. 305–307.

- [187] WU, P., LIU, X., AND LIU, J. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia* 25 (2022), 1674–1685.
- [188] YOUNGBLOOD, M. Extremist ideology as a complex contagion: the spread of far-right radicalization in the united states between 2005 and 2017. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–10.
- [189] YUMATLE, C. Pluralism. *The Encyclopedia of Political Thought, First Edition. Edited by Michael T. Gibbons, Published* (2015).
- [190] ZHANG, C., LI, Z., AND ZHANG, K. An inverse propensity score framework for breaking filter bubble recommendation. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)* (2023), IEEE, pp. 1588–1592.