DEEPFAKES GUIDELINES

Version 1

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Contents

1. Executive Summary

The rapid advancement of Artificial Intelligence (AI) technologies, particularly deepfakes, has brought significant opportunities and challenges across various sectors. Deepfakes, hyper-realistic synthetic media created using deep learning techniques, have the potential to manipulate audio, video, or other digital content in ways that are difficult to distinguish from reality. While deepfakes can be used beneficially in fields such as marketing, entertainment, retail, education, healthcare, and cultural applications, they also pose severe risks, including identity fraud, non-consensual manipulation, and the spread of disinformation.

This document provides comprehensive guidelines to address the implications of deepfake technologies and mitigate their associated risks. It defines deepfakes and their impact on society, raising awareness about both their malicious and non-malicious applications. The guidelines emphasize the importance of ethical principles in developing and using deepfake technologies, including privacy, transparency, accountability, and social benefits.

For deepfake technology developers and content creators, the recommendations include implementing strong data protection measures, securing consent for using personal data, and maintaining transparency by providing clear documentation and explanations of how deepfakes are generated. Developers are also urged to ensure accountability by establishing human oversight and clear responsibilities for AI outcomes. Furthermore, the guidelines highlight the importance of directing deepfake technology towards applications that provide social and environmental benefits. Adhering to these principles not only protects individuals and society from potential harms but also fosters trust and acceptance of AI technologies.

Consumers of deepfake technology are advised to adopt best practices for detecting and responding to deepfakes. This involves verifying the authenticity of the source, analyzing audio-visual elements for inconsistencies, and utilizing AI tools to detect signs of manipulation. Public awareness and education campaigns are also crucial to help individuals recognize and respond to deepfakes effectively. These measures are vital for maintaining the integrity of information and protecting individuals from the malicious use of deepfake technology.

The guidelines also highlight the necessity of continuous learning and skill development to manage AI applications effectively. Workshops and hands-on training sessions are recommended to equip individuals and organizations with the knowledge needed to handle AI technologies. Enhancing organizational capacity and capabilities through tailored workshops and strategic hiring practices can significantly improve innovation and problem-solving capabilities, enabling entities to navigate the complexities of AI and deepfake technologies. Emphasizing the importance of ongoing education ensures that both individuals and organizations remain adept at handling emerging challenges and leveraging new opportunities.

By following these comprehensive guidelines, stakeholders can harness the positive potential of deepfakes while minimizing their risks. This approach ensures that deepfake technology is used ethically and responsibly, fostering innovation and maintaining public trust in AI advancements.

2. Introduction and Problem Statement

Artificial Intelligence (AI) technologies have increasingly become integral parts across sectors, significantly impacting the global economy. These technologies support a wide range of applications leading to opportunities in industries such as healthcare, finance, transportation, and entertainment. However, these opportunities come with significant risks. AI, in general, can be misused in various ways, leading to risks such as bias and privacy invasion. Particularly with the advent of deepfake technologies, these risks are magnified. Deepfakes are hyper-realistic synthetic media created using deep learning techniques that manipulate audio, video, or other digital content, making it difficult to distinguish between real and fake content.

While deepfakes have legitimate uses, they pose severe risks, including the potential for disinformation, fraud, and erosion of trust in media.

Addressing the issues posed by deepfakes is crucial to maintaining the integrity of information, protecting individuals' reputation, and ensuring public safety. The difficulty in detecting and identifying deepfakes exacerbates these risks, making it imperative to develop comprehensive policy and regulatory instruments as well as robust countermeasures.

As shown in figure 1, deepfakes are one of the several use cases of generative AI, utilizing technologies such as Generative Adversarial Networks (GANs) and neural networks (for further details refer to the GenAI guidelines for the general public).

GENERATIVE AI USE CASES

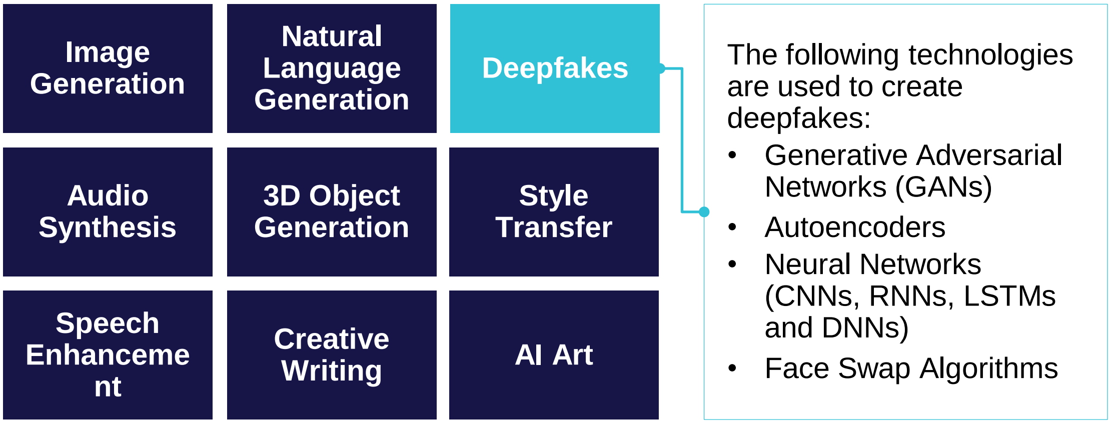

Image Generation	Natural Language Generation	Deepfakes
Audio Synthesis	3D Object Generation	Style Transfer
Speech Enhanceme nt	Creative Writing	AI Art

The following technologies are used to create deepfakes:

- Generative Adversarial Networks (GANs)
- Autoencoders
- Neural Networks (CNNs, RNNs, LSTMs and DNNs)
- Face Swap Algorithms

Figure 1: Generative AI Use Cases .

Unlike traditional image and audio editing tools, deepfakes leverage advanced algorithms to create highly realistic altered content. This technology enables the generation of fake media that is almost indistinguishable from real footage, posing significant challenges for detection and verification (Figure 2).

[1]IBM, Norton

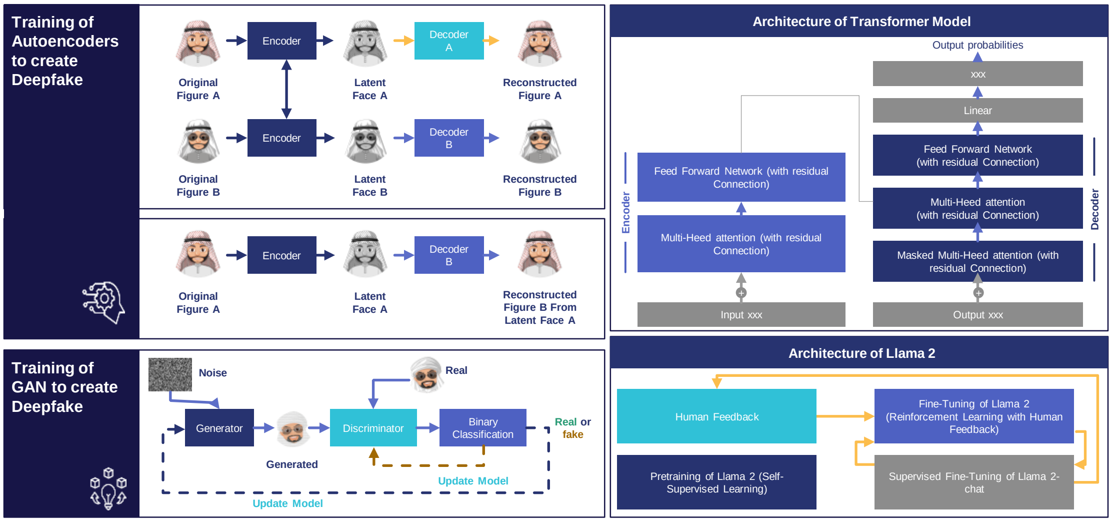
Training of Autoencoders to create Deepfake
Original Figure A
Encoder
Latent Face A
Decoder A
Reconstructed Figure A
Original Figure B
Encoder
Latent Face B
Decoder B
Reconstructed Figure B
Original Figure A
Encoder
Latent Face A
Decoder B
Reconstructed Figure B From Latent Face A
Architecture of Transformer Model
Output probabilities
xxx
Linear
Feed Forward Network (with residual Connection)
Multi-Heed attention (with residual Connection)
Masked Multi-Heed attention (with residual Connection)
Output xxx
Feed Forward Network (with residual Connection)
Multi-Heed attention (with residual Connection)
Input xxx
Encoder
Decoder
Training of GAN to create Deepfake
Noise
Real
Generator
Generated
Discriminator
Binary Classification
Real or fake
Update Model
Update Model
Architecture of Llama 2
Human Feedback
Fine-Tuning of Llama 2 (Reinforcement Learning with Human Feedback)
Pretraining of Llama 2 (Self-Supervised Learning)
Supervised Fine-Tuning of Llama 2-chat

Figure 2: Processes for generating deepfakes .

Deepfake technology originated from GANs in 2014, enabling AI to generate new data from learned patterns. Initially requiring extensive data and computational power, deepfakes were limited to skilled professionals. Advances in AI and software have since made deepfakes more accessible, affordable, and faster to create. As AI technology progresses, deepfakes will become more realistic and harder to detect, necessitating continuous improvements in detection methods and regulatory measures to mitigate associated risks.

Deepfakes are not inherently malicious; their intent and application determine their impact (Figure 3).

APPLICATIONS OF DEEPFAKES

🟩 Non-Malicious Deepfakes 🟥 Malicious Deepfakes

Image Generation Manipulate photos using AI to alter or generate realistic images of people or scenes	🟩 **Training Simulations** Creating realistic simulations for training purposes in fields such as military, medical, or customer service	🟥 **Explicit Material** Attackers use deepfakes to generate explicit images to harass or blackmail individuals	🟩 **Advertising** Brands use deepfakes to create engaging and novel advertisements	...
Audio Generation Synthetic audio recording mimicking a person's voice used to create realistic but fake speech	🟥 **Financial Scams** Using audio to mimic voices of trusted people to trick victims into transferring money	🟩 **Audio Deepfakes Enable** Realistic voiceovers and dubbing in different languages without the original actors	🟥 **Voice Cloning** Using cloned voices to access bank information to authorize transactions or commands	...
Video Generation AI-generated video that alters the appearance of individuals to create realistic but fake visual content	🟩 **Film and Production** Similar to image deepfakes, video deepfakes are used to create digital actors, de-age performers, or even alter dialogue	🟥 **Political Manipulation** Using deepfake technology, people can altering interviews and speeches to discredit candidates	🟥 **Corporate Espionage** Using deepfakes to fake executive communications to steal sensitive information	...

Other forms of digital content

As of June 2024, **no documentation** explicitly reference deepfakes in **other forms of digital content**. However, potential **future examples** include **genetic deepfakes**, **olfactory deepfakes**, and **taste-based deepfakes**

Figure 3: Applications of deepfakes technology .

Malicious deepfakes are AI-generated media designed to deceive, harm, or exploit individuals, often used for spreading false information, defamation, or fraud (Figure 4).

[2] IBM, Norton
[3] The Medium

	Malicious Deepfakes	Non-Malicious Deepfakes
	Definition: ***AI-generated media designed to deceive, harm, or exploit individuals, often used in spreading false information, defamation, or fraud***	**Definition:** ***AI-generated media created for benign purposes like entertainment, satire, or education, without the intent to deceive or cause harm***
Transparency & Disclosure	Deepfake content without disclosing its synthetic nature. This lack of transparency can manipulate public perception and cause harm	Clearly labelled content to inform viewers that it is synthetically generated or altered. This transparency ensures that audiences are aware they are seeing manipulated media
Authorization & Consent	Using individuals' likeness without their consent. This violates their rights and privacy. Unauthorized use can result in identity theft and defamation	Obtain explicit consent from individuals whose likeness is used in the creation of deepfakes. This respects personal rights and prevents misuse of identities
Application	Employ deepfakes for harmful purposes such as spreading false information, political manipulation, or defamation	Use deepfakes for constructive purposes like education, entertainment, historical reconstructions, and accessibility.
Compliance	Failing to follow ethical guidelines and regulations in the creation and use of deepfakes. The absence of accountability and oversight can result in widespread harm	Adherence to comprehensive ethical guidelines for the creation and use of deepfakes through independent monitoring to maintain accountability and respect

3. Purpose and Scope of the Guidelines

Figure 4: Differentiating factors between malicious and, non-malicious deepfakes .

This document provides a thorough explanation of deepfake technologies, their implications, and the necessary measures to mitigate their risks for technology developers, content creators, and consumers.

The guidelines focus on:

- Defining deepfakes and their impact on society
- Raising awareness about the malicious and non-malicious applications of deepfakes technology
- **Establishing ethical principles for Deepfake Technology Developers**: Ethical guidelines are laid out for developers to promote responsible creation and implementation of deepfake technologies, emphasizing transparency, consent, and respect for privacy.
- **Defining clear guidance for Deepfake Content Creators**: This includes specific instructions for content creators on how to responsibly produce and distribute deepfake content, ensuring they adhere to ethical standards and legal requirements.
- **Providing clear guidance for Deepfake Consumers**: The guidelines aim to educate and raise awareness about the potential risks and benefits of deepfake technologies. They include measures for identifying and protecting against malicious deepfakes and advice on how to responsibly engage with and share deepfake content.
- Providing risk mitigating, and enabling guidance for regulators and government entities
- Presenting protection measures to identify and counteract deepfakes
- Encouraging ongoing education and awareness to build resilience against deepfake-related threats.

This document is applicable to three primary groups involved with deepfake technologies: technology developers, content creators, and consumers.

[4] The Prindle Institute for Ethics, Walton College

4. Overview of Malicious Deepfakes

As deepfake technology becomes more sophisticated and accessible, the associated risks have increased globally, creating a profound sense of vulnerability among citizens. Advanced identity fraud schemes, enabled by these tools, are increasingly targeting individuals and organizations, leading to significant security and privacy concerns. The rapid increase in deepfake-specific fraud cases underscores the urgency for global cooperation in combating these sophisticated threats.

The following sections detail the primary risk categories associated with malicious deepfakes:

4.1. Imposter Scams

Imposter scams use deepfake technology to convincingly impersonate trusted individuals to extract sensitive information or manipulate victims. By mimicking voices, facial expressions, and mannerisms, these scams exploit the inherent trust placed in the impersonated person. Common scenarios include fake calls or video conferences where the imposter, posing as a known executive or authority figure, requests confidential information or authorizes financial transactions. The sophistication of these scams challenges traditional security measures, necessitating advanced detection techniques and heightened awareness to prevent fraud.

Example:

Multinational Firm Scam : In a significant case reported by authorities in an Asian region, an employee at a multinational firm was tricked into paying out a substantial amount of money to fraudsters using deepfake technology to impersonate a senior executive during a video conference call. This scam demonstrates the potential for deepfakes to facilitate large-scale financial fraud.

4.2. Non-consensual Manipulation

Non-consensual manipulation uses deepfake technology to create explicit or compromising content of individuals without their consent, often for harassment, blackmail, or reputational damage. Victims face severe emotional distress and potential long-term mental health issues, strained relationships, and professional setbacks. This violation of privacy and trust underscores the urgent need for robust legal and technological safeguards.

Example:

Celebrity Deepfake Scandal : A few years ago, explicit deepfake images and videos featuring several celebrities were widely circulated online without their consent. This non-consensual use of deepfake technology led to significant emotional distress and privacy violations for the affected individuals.

[5] CNN
[6] Yahoo News

4.3. Disinformation and Propaganda

Deepfakes can spread false information, manipulate public perception, and influence political outcomes by depicting political figures making false statements or engaging in unethical behavior. This tactic can undermine opponents, sway public opinion, and destabilize societies, leading to widespread misinformation, eroded trust in public institutions and social unrest. The convincing nature of deepfakes poses a significant threat to political discourse and societal stability.

4.4. Example:

Political Leader Deepfake Demonstration : In a recent election, deepfakes were used to create fake Non-consensual Manipulation

Non-consensual manipulation uses deepfake technology to create explicit or compromising content of individuals without their consent, often for harassment, blackmail, or reputational damage. Victims face severe emotional distress and potential long-term mental health issues, strained relationships, and professional setbacks. This violation of privacy and trust underscores the urgent need for robust legal and technological safeguards.

Example:

Celebrity Deepfake Scandal : A few years ago, explicit deepfake images and videos featuring several celebrities were widely circulated online without their consent. This non-consensual use of deepfake technology led to significant emotional distress and privacy violations for the affected individuals.

[7] Business Insider
[8] Yahoo News
[9] Business Insider

5. Guidance for Deepfake Technology Developers

This section outlines the importance of AI ethics in deepfake technology development.

Overall, developers should ensure deepfake technology adheres to ethical standards, safeguards personal data, and prevents misuse. Deepfake technology must comply with laws, respect social ethics, and align with correct values. Technology should enable content authenticity verification and prevent the misuse of synthetic media, promoting responsible and trustworthy use.

This is paramount due to the profound societal impact of deepfakes and help steer the development of deepfakes towards positive and constructive uses while mitigating associated risks.

Developers should thus adhere to all the points in the below list when developing and managing deepfake technologies.

1. Regulatory Compliance

1.1. Adhere to all relevant local and international data privacy laws, such as GDPR, CCPA, and KSA's PDPL and Anti-Cyber Crime Law, and integrate legal compliance checks into your development environment (e.g., policy-as-code frameworks such as Terraform and Cedar Policy Language) (PDPL Article 3, AI Ethics Principles – Compliance Section, page 32)

2. Data Privacy and Protection

2.1. Implement strict data collection protocols, using only the minimal necessary data for intended purposes, and applying anonymization and pseudonymization to reduce re-identification risks.(PDPL Article 19, AI Ethics Principles – Principle 2)

2.2. Implement strict data collection protocols, using only the minimal necessary data for intended purposes, and applying anonymization and pseudonymization to reduce re-identification risks. (PDPL Articles 10, 11, 12, 14, and 18; AI Ethics Principles – Principle 2)

2.3. Integrate privacy-preserving techniques into data processing, minimize data retention, and establish automated review cycles to regularly assess and securely delete unnecessary data (e.g., using cryptographic erasure), reducing the risks of breaches and unauthorized access. (PDPL Articles 4 and 28; AI Ethics Principles – Principles 2)

2.4. Implement consent management systems integrated into your AI tools, ensuring that data used in training has been properly consented to. Securely record and manage consent transactions. (PDPL Articles 4 and 28; AI Ethics Principles – Principles 2)

3. Transparency and Explainability

3.1. Document every aspect of your AI model, including data sources, preprocessing steps, algorithmic architecture, and the decision-making process. Ensure this documentation is machine-readable and integrated into your CI/CD pipeline for continuous updates. (AI Ethics Principles – Principles 5 and 7)

3.2. Incorporate explainability features within AI models using techniques such as LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) to ensure that users and stakeholders can understand how outputs are generated. (AI Ethics Principles – Principle 6)

[8] Advanced Encryption Standard (AES) is a symmetric block cipher that uses the 256-bit key length to encrypt as well as decrypt a block of messages.

[9] A series of automated steps that helps software teams deliver code faster, safer, and more reliably.

[10] A technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction.

[11] SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model.

3.3. Develop and implement non-intrusive digital watermarking techniques (e.g., steganographic methods) and ensure these markers are resistant to common attacks (e.g., re-compression, cropping, and filtering) (AI Ethics Principles – Principle 2)

3.4. Maintain transparency by regularly publishing reports on the use and impact of your deepfake technology, including measures taken to prevent misuse. (AI Ethics Principles – Principle 7)

4. Bias

4.1. Implement processes like sampling and data augmentation to identify and eliminate biases in the data, and maintain comprehensive documentation of the data sources, preparation steps, and validation processes. (PDPDL Article 14; AI Ethics Principles – Principle 1)

5. Accountability and Responsibility

5.1. Integrate Human-in-the-Loop (HITL) mechanisms for continuous oversight during key stages like training, validation, and deployment to ensure ethical decision-making and mitigate risks associated with autonomous model behavior. (AI Ethics Principles – Principle 7)

5.2. Develop and implement a comprehensive governance framework that includes role-based access controls (RBAC), version control systems for models, and audit trails for all model iterations. (AI Ethics Principles – Principle 7)

5.3. Set up an automated reporting system that uses log analysis and machine learning to detect and report any unauthorized or unethical use of deepfake technology.

6. Social Responsibility

6.1. Ensure that deepfake technology towards socially and environmentally beneficial applications, such as in education, entertainment, and healthcare. Misuse or unethical application may result in legal consequences. (Article 8; AI Ethics Principles – Principle 4, Anti-Cyber Crime Law Article 3)

6.2. Develop AI algorithms (e.g., such as Amazon SageMaker, ELKI, PyOD) to detect and flag suspicious activity patterns, such as rapid media dissemination or unusual engagement spikes, as cybercrimes may result in imprisonment, fines, or both. Implement features to identify and restrict inauthentic accounts and behavior using machine learning techniques (e.g., anomaly detection, behavior analysis). (Anti-Cyber Crime Law Article 3)

To ensure that the development and application of deepfake technology aligns with the highest ethical standards, a risk assessment framework needs to be incorporated throughout the different stages from development to deployment.

This framework is designed to address potential risks and enforce adherence to the key ethical principles outline.

The detailed risk assessment framework, which outlines specific questions and mitigation strategies for each phase, can be found in the appendix.

6. Guidance for Deepfake Content Creators

Content creators must ensure that deepfake technology is used responsibly, adhering to ethical standards, safeguarding personal data, and preventing misuse. Compliance with laws, respect for social ethics, and alignment with correct values are imperative.

Given the profound societal impact of deepfakes, it is essential to steer content created towards positive and constructive outputs.

The following requirements are designed to maintain public trust, prevent misuse, and address risks such as imposter scams, non-consensual manipulation, disinformation, and propaganda.

1. Regulatory Compliance

1.1. Deepfake services must not be used to produce, reproduce, publish, or transmit information that is prohibited by laws or administrative regulations. Content creators must ensure all deepfake content complies with relevant data protection, intellectual property, and industry-specific regulations. (PDPL Article 5; Anti-Cyber Crime Law Article 6; AI Ethics Principles – Principle 6)

2. Data Privacy and Protection

2.1. Implement robust security protocols, including end-to-end encryption, multi-factor authentication (MFA), and strict access control mechanisms, to protect personal data involved in deepfake content creation. Ensure that data storage systems are resilient to breaches and unauthorized access. (PDPL Article 19, AI Ethics Principles – Principle 2)

2.2. Develop and maintain an incident response plan to quickly address and mitigate any data breaches or security threats, ensuring minimal impact on personal data integrity. (AI Ethics Principles – Principle 2)

2.3. Secure explicit, written consent from individuals or entities prior to using their data in deepfake content creation. (PDPL Articles 4, 7, and 11; AI Ethics Principles – Principle 7)

2.4. Maintain consent records securely, ensuring they are encrypted and backed up to prevent loss or tampering. Consent management systems should be auditable and compliant with relevant data protection laws. (PDPL Articles 4, 7, and 11; AI Ethics Principles – Principle 2)

2.5. Provide data subjects with detailed information about how their data will be used, including specific purposes, potential risks, and the broader context of the content's application. (PDPL Articles 4, 7, and 11; AI Ethics Principles – Principle 2)

2.6. Store consent records in a secure, accessible format, ensuring they are easy to retrieve for audits or legal verification. (PDPL Articles 4, 7, and 11; AI Ethics Principles – Principle 2) A sample consent form can be found in the appendix.

2.7. Ensure that personal data is used exclusively for the purposes specified in the consent agreement. Any expansion of the data's use beyond the original consent requires a new, explicit agreement. (PDPL Articles 4, 16 and 18; AI Ethics Principles – Principle 2)

2.8. Clearly communicate the right to withdraw consent at any time, with a straightforward process in place for individuals to exercise this right. Upon withdrawal, ensure that the data is no longer used and is securely deleted if requested. (PDPL Article 5)

3. Transparency and Explainability

3.1. Maintain detailed records of the deepfake creation process, including data sources, algorithms, tools, and techniques used. Ensure this documentation is version-controlled and includes the rationale behind methodological choices. (AI Ethics Principles – Principle 6)

3.2. Utilize automated systems for applying watermarks to ensure consistency and prevent accidental omission. These systems should be integrated into the content creation workflow to minimize human error. (AI Ethics Principles – Principle 6)

3.3. Implement visible, tamper-resistant watermarks (e.g., using steganographic methods) on all deepfake content, ensuring that the synthetic nature of the content is always identifiable, regardless of the platform or medium. Watermarks should be designed to be durable across different formats and media types. (AI Ethics Principles – Principle 6)

3.4. Implement robust version control systems for all documentation, ensuring all changes are tracked, recorded, and easily traceable.

4. Accountability and Responsibility

4.1. Take responsibility for the content generated using deepfake technology and ensure it complies with relevant regulations and standards. (AI Ethics Principles – Principle 7)

4.2. Develop stringent internal guidelines that cover the ethical, legal, and technical standards for deepfake content creation. Ensure these guidelines are regularly updated to reflect best practices and industry standards. (AI Ethics Principles – Principle 7)

4.3. Ensure deepfake content is distributed through secure, controlled channels to prevent unauthorized access or alteration. Implement DRM and secure CDNs to protect content integrity. Violations may lead to penalties, including fines or imprisonment. (AI Ethics Principles – Principle 5; Anti-Cyber Crime Law Article 3)

4.4. Integrate cryptographic hashing and blockchain technology to create immutable records of the original content. These tools should enable the verification of content integrity, ensuring that any alterations are detected and traced back to the source. (AI Ethics Principles – Principle 7)

7. Guidance for Regulators

Effective regulation of deepfake technology is crucial to safeguard against privacy and security risks. Regulators can enforce a range of measures to minimize risks, however, regulation should also focus on unlocking the benefits of deepfakes.
This section aims to ensure that deepfakes are regulated while enabling potential benefits

1. Platform Regulation and Monitoring

1.1. Require internet intermediaries, mass media service providers, and social media platforms to promptly disable access to or prevent the dissemination of deepfake content identified as false or misleading. This includes establishing a clear process for reporting and removing such content.

1.2. Regularly monitor these platforms to ensure compliance with directives and impose penalties for failure to act on identified harmful content.

2. Risk Assessment and Approval Processes

2.1. Conduct thorough assessments of the deepfake technologies being used or approved for use within the country. This includes reviewing the algorithms, data sources, and ethical considerations in the development of such technologies.

2.2. Develop a formal approval process for deepfake technologies, ensuring they meet established ethical, legal, and technical standards before being deployed or used commercially.

2.3. Develop specific risk assessment frameworks for government entities that will use or approve deepfake technologies.

2.4. Develop and enforce set standards for producing generative content, such as those outlined by the Coalition for Content Provenance and Authenticity (C2PA), to ensure transparency and accountability in the creation and dissemination of deepfake and other AI-generated media.

3. Penalties and Sanctions

3.1. Ensure that penalties and sanctions for the misuse of deepfake technology are proportional to the severity of the offense. Consider scaling penalties based on factors such as intent, impact, and recurrence.

3.2. Introduce provisions that limit penalties or sanctions for de minimis or incidental uses of deepfakes. *(For example, if a movie briefly shows a crowd scene where a digital replica of a well-known person is included among many others, this would be considered minimal and not central to the scene, and thus excluded from penalty considerations.)*

4. Transparency and Accountability

4.1. Require regular reporting (e.g., on an annual basis) on the use of deepfake technology by government entities, including the purposes, outcomes, and any issues encountered. This will help maintain transparency and public trust.

4.2. Implement regular audits and oversight mechanisms to ensure that deepfake technology is being used ethically and in line with regulatory frameworks. Audits should be conducted by independent bodies to ensure objectivity.

4.3. Develop and maintain an annual Use Case Inventory that collects data from various government agencies on deepfake technology usage. This inventory should include the total content produced, and trends over time.

5. Education and Public Awareness

5.1. Develop and implement mandatory training programs for government employees and regulators on the ethical, legal, and technical aspects of deepfake technology. This training should be updated regularly to reflect new developments and challenges in the field.

5.2. Collaborate with organizations like EndTAB (Ending Tech-Enabled Abuse), which provides resources for education and reporting abuse. Leveraging their expertise can enhance the effectiveness of training programs and public awareness campaigns, ensuring that both government employees and the public are better equipped to handle and report tech-enabled abuses, including deepfakes.

5.3. Launch public awareness campaigns to educate citizens about deepfake technology, its potential uses, and the safeguards in place to protect against misuse. This will help in fostering an informed public discourse.

5.4. Implement annual reporting on deepfake-related education programs for government employees. These reports should include detailed information on the number of training sessions conducted, the number of attendees, the content covered, and the overall effectiveness of the programs.

5.5. Mandate that all relevant government employees participate in recurring annual training or refresher workshops on deepfakes to ensure they stay informed about the latest developments and challenges in the field.

8. Guidance for Deepfake Consumers to Detect Deepfakes and Prevent Risks

Identifying deepfakes can be effectively achieved through a three-step approach to scrutinize content authenticity. Below is a detailed method to detect potential deepfakes:

8.1. Assess the Message:

Source: Verify if the source is trustworthy, such as official organizations or known individuals. Cross-check with official websites or contact the source directly through known channels to ensure authenticity.

Context: Evaluate if the content aligns with what is expected from the source. For example, public officials typically do not endorse investment schemes or promote unusual offers.

Aim of Content: Be wary of messages that demand urgent action, such as clicking on links, downloading apps, or providing personal information. Discuss suspicious messages with friends or family to get their perspective.

8.2. Analyze Audio-Visual Elements:

Irregular facial movements: Look for unnatural or awkward facial expressions, especially around the mouth, eyes, and eyebrows. Deepfakes may struggle with subtle movements like micro-expressions or slight muscle twitches. Although advanced deepfakes have improved, these small irregularities can still indicate manipulation.

Lighting inconsistencies: Despite advancements, deepfakes may still show lighting inconsistencies. Look for shadows or highlights on the face that don't align with the light source or surroundings. Such discrepancies, especially in complex or dynamic scenes, can indicate manipulation.

Skin color changes: Skin tone changes are less noticeable with advanced rendering, but irregularities can still appear during lighting transitions or quick movements. Watch for sudden shifts in skin texture or tone that don't match the scene, as these may indicate manipulation.

Blinking patterns: Modern deepfakes can simulate blinking more naturally, but the subtleties of eye movement can still give away a fake animation. Look for unnatural eye movements, such as eyes that seem too fixed, or blinking that seems either too slow, too fast, or inconsistent with the emotional context of the scene. Subtle asynchrony between blinks and the emotional state of the subject can still be a clue.

Lip-sync issues: Lip-sync technology has advanced significantly, making mismatches harder to detect. However, slight delays, unnatural jaw movements, or lips that don't fully close during certain sounds can still reveal inconsistencies. Pay attention to the synchronization of lip movements with complex phonemes, where errors may be more noticeable.

High-Resolution Anomalies: As deepfakes improve in visual fidelity, irregularities may become visible only under close scrutiny or at high resolutions. Zoom in on high-resolution footage to detect subtle artifacts around the edges of the face, particularly where the face meets the hairline, ears, or background. Artifacts or unnatural blending in these areas can still indicate deepfake usage.

Dynamic Scene Analysis: Deepfakes often struggle with rapid movements or complex scenes involving multiple lighting sources or intricate backgrounds. Analyze scenes where the subject moves quickly or interacts with complex environments. Look for motion blur inconsistencies or unnatural blending of the subject with the background during dynamic sequences.

8.3. Authenticate Content Using Tools:

Detection Tools: Utilize AI tools that analyze content for signs of manipulation. These tools can examine various aspects of media to detect inconsistencies that might indicate deepfake technology:

- ***Pixel Analysis:*** This technique involves scrutinizing the pixel-level details of an image or video to uncover anomalies that could suggest manipulation. Such analysis can detect inconsistencies in lighting, shadows, and facial expressions, which are often telltale signs of deepfakes.
- ***Motion Inconsistencies:*** Tools like motion analysis software can identify irregularities in the way objects or people move within a video. Since deepfakes often struggle to replicate natural movements perfectly, these tools can help detect unnatural motion patterns.
- ***Audio-Visual Synchronization:*** Some tools are designed to assess whether the audio matches the visual content. Deepfakes might struggle with synchronizing lip movements with spoken words, especially in cases where the audio has been artificially generated or manipulated.
- ***AI-Based Analysis:*** Open-source tools such as Deepware Scanner and Sensity AI provide consumers with accessible methods to detect manipulated media. These tools use AI to analyze videos for common deepfake indicators, such as unnatural facial movements or artifacts left by manipulation.
- ***Disinformation Detection:*** Tools such as the Global Disinformation Index perform cross-platform open-source intelligence tracking of disinformation online and can also be used to detect deepfakes.

Content Provenance: Establishing the origin of content is crucial in verifying its authenticity. By examining metadata and embedded digital watermarks, it is possible to trace the content back to its source. Here are some methods and tools used for content provenance:

- ***Metadata Analysis:*** Metadata contains information about the file, such as the date and time it was created, the device used, and the software employed. By analyzing this data, it’s possible to identify discrepancies or signs of tampering. For example, inconsistencies between the creation date of the file and the supposed event it depicts can be a red flag.

- ***Digital Watermarking:*** Watermarks are embedded into the content to provide proof of origin. Advances in technology have enabled the embedding of imperceptible digital watermarks within images and videos that can be traced back to the content creator. These watermarks are resistant to common editing techniques and can be used to verify whether content has been altered.
- ***Content Authenticity Tools:*** Tools like the Adobe Content Authenticity Initiative or Truepic offer solutions that ensure the integrity of digital content. These platforms embed content credentials at the point of capture, enabling users to trace the history of the media and verify that it hasn't been tampered with.
- ***Blockchain Technology:*** Blockchain is being explored as a method to track and authenticate content. By storing content on a decentralized ledger, blockchain can provide an immutable record of the content's origin and any modifications it has undergone. This technology can be used to create a transparent and traceable history of the content, making it more difficult to pass off deepfakes as genuine.

Leveraging Industry Tools Standards: As the use of deepfakes grows, industry leaders are developing and adopting standards to ensure content authenticity. These standards often involve the use of tools and platforms that can detect manipulations and authenticate content. Examples include:

- ***YouTube Content ID:*** YouTube's Content ID system is designed to identify copyrighted material uploaded to the platform. While not specifically designed for deepfakes, the system can be adapted to recognize manipulated content by analyzing the audio and video tracks.
- ***Adobe's Project Origin:*** Adobe's Project Origin works by attaching a verifiable origin to media, ensuring that viewers can trust the source of the content.

8.4.Report the Incident:

If someone is impacted by a deepfake, specific procedures can be followed to mitigate the damage and protect their reputation. Below is a step-by-step process intended to provide a structured approach for victims of deepfake incidents.

However, it is important to note that these guidelines operate within the broader legal framework of the Kingdom of Saudi Arabia. Where existing laws or administrative regulations provide more specific provisions, those will take precedence over the guidance provided here. Therefore, the overarching laws and regulations of the Kingdom should always be the primary reference for addressing deepfake incidents:

- ***Documenting Evidence:*** Victims should immediately save copies of the deepfake content, including links, screenshots, and metadata. It is recommended to compile a detailed file with all relevant evidence, including the dates, times, and platforms where the deepfake was shared, ensuring this documentation is ready for legal proceedings or reporting to authorities.

- ***Identify type of Malicious Use:*** It is important to accurately determine whether the deepfake is being used for fraud, defamation, blackmail, or non-consensual dissemination of personal data. The victim should categorize the deepfake by its intended malicious use and prepare to report accordingly, ensuring that each aspect of the misuse is thoroughly documented to strengthen any legal actions. For support in identifying the specific type of deepfake, victims may also refer to platforms like the Media Manipulation Casebook that provide detailed examples and research on various forms of media manipulation. Under Article 36 of the Personal Data Protection Law, violations will be examined, and warnings will be issued, or fines will be imposed considering the type of violation committed, its seriousness and the extent of its impact. Further, under Article 24 of the Personal Data Protection Law, disclosure of personal data without consent is prohibited.
- ***Reporting the Incident to the Platform:*** The deepfake should be reported immediately to the platform where it was shared (e.g., Twitter, Facebook), requesting its removal. When reporting, the victim should aim to reference the platform's Terms of Service, especially clauses that prohibit the violation of user privacy and the dissemination of harmful content. Most social media platforms have strict policies against content that infringes on privacy rights, and citing these specific terms can strengthen the case for prompt action. Additionally, victims can seek support from organizations like the Cyber Smile Foundation that are dedicated to combating all forms of online abuse, including deepfake related incidents.
- ***Reporting the Incident to Local Authorities:*** After reporting the deepfake to the platform, victims should also notify relevant local authorities to ensure further action. This can for example be done by re-sending the malicious message to 330330 or submitting a complaint via the Kollona Amn (كلنا أمن) app, which directly connects users to law enforcement. Additionally, victims should report the incident to the Cybercrime Unit within the Ministry of Interior through their online portal or by visiting a police station. For cases involving financial fraud or identity theft, reporting to the Saudi Central Bank's e-service and their bank is crucial to prevent further damage and initiate protective measures. Article15 of the Anti-Cyber Crime Law stipulates that the Bureau of Investigation and Public Prosecution is responsible for investigating cybercrimes. Article 14 of the Anti-Cyber Crime Law stipulates that CST shall provide assistance and technical support to competent security agencies during the investigation stages of cybercrimes.
- ***Seeking Legal Advice:*** It is essential to consult with a lawyer experienced in digital rights and data protection laws. Victims should engage a legal professional to guide the process, providing them with all evidence and legal bases to prepare for both civil and criminal proceedings.
- ***Employing Digital Forensics:*** Engaging digital forensics experts to trace the origins of the deepfake and utilizing AI detection tools is advisable. It is also recommended to contract a reputable digital forensics firm to analyze the deepfake content, ensuring the findings are documented and can be effectively used.
- ***Enhancing Personal Security:*** Victims are advised to implement enhanced security protocols across all digital platforms and consider subscribing to monitoring services to safeguard against further misuse of personal data. Strengthening privacy settings on all online accounts, enabling two-factor authentication, and regularly monitoring for further incidents are critical steps.
- ***Refer to Overarching Laws:*** This document provides guidance on the development, use, and consumption of deepfakes. It is important to note that where laws or administrative regulations apply, those provisions should take precedence. As such, leveraging the PDPL, Anti-Cyber Crime Law, and Criminal Code is essential for addressing deepfake incidents. Victims should work with their lawyer to develop a comprehensive legal strategy that references these specific laws to pursue all available legal options in response to the deepfake incident.

8.5. Best practices for protecting oneself and others:

To minimize the risks associated with deepfakes, individuals and organizations can adopt several best practices:

- **Limiting Personal Data Exposure Online:** Reduce the amount of personal data and images shared on social media and other public platforms to limit the raw material available for creating deepfakes. Avoid posting high-resolution images and personal videos that can be exploited for malicious purposes.
- **Educating the Public on Deepfake Risks:** Raise awareness about the existence and dangers of deepfakes. Inform people on how to recognize and respond to suspicious content. Public education campaigns can include workshops, online resources, and collaboration with media outlets to disseminate information about deepfakes.
- **Learning About Deepfake Characteristics:** For individuals, it is crucial to learn more about the characteristics of deepfakes. Understanding the typical signs of deepfake content, such as unnatural facial movements, irregularities in lighting and shadows, and inconsistencies in audio, can help in identifying and questioning the authenticity of such media. Staying informed through reputable sources and utilizing available tools to analyze and verify content can significantly reduce the impact of deepfake threats.
- **Implementing Robust Verification Processes:** For organizations, especially those in media and communication, implementing strict verification processes for the content they disseminate can help ensure authenticity. This includes using AI detection tools as well as manual verification methods by trained professionals. These verification processes could include:
  - Using AI Tools and Checking Metadata: Utilize advanced AI tools to analyze videos and images for signs of manipulation. Additionally, check the metadata of digital content, which provides information about the creation date, editing history, and software used, helping to verify if the content has been tampered with.
  - Manual and Expert Verification:
    - Fact-Checking Services: Utilize professional fact-checking services like to confirm the authenticity of suspicious content.
    - Expert Analysis: Digital forensics experts can manually review content for signs of deepfake manipulation, such as irregular facial features, unnatural movements, and mismatches between audio and visuals.
  - Cross-Referencing Sources and Watermarking: Compare the suspicious content with trusted and official sources. If a video or audio clip seems dubious, check it against known authentic versions or contact the source directly to confirm its validity. Additionally, protect original content with digital watermarks and signatures. These technologies help verify that media has not been altered since its creation.
  - Promoting Skills Development: Encourage continuous learning and skills development in AI and cybersecurity fields. Providing training programs and workshops for employees and the public can enhance the ability to detect and respond to deepfake threats effectively.

9. Overview of Non-Malicious Deepfakes

Deepfake technology, while often associated with malicious uses, also possesses transformational potential across various sectors when applied ethically and responsibly. Figure 5 highlights six main sectors where non-malicious applications of deepfakes can create opportunities in six sectors: **Marketing, Entertainment, Retail, Education, Healthcare, and Culture.**

Sector	Description	Example
Marketing	**Support Saudi SMEs** create **personalized and engaging advertisements**, enhancing marketing efforts.	The **Ministry of Commerce** with **SDAIA** can support Saudi businesses create virtual influencers to promote products on social media, offering personalized marketing at scale
Entertainment	**Enhance Saudi's entertainment industry** by enabling the **creation of digital characters** and **reducing production costs**	The **General Entertainment Authority** with **SDAIA** can use deepfakes to create virtual hosts for events and TV shows, enhancing viewer engagement
Retail	**Transform the retail sector** in Saudi Arabia by enabling **virtual try-ons**, enhancing **customer engagement**, and providing **personalized shopping experiences**	The **Ministry of Commerce** with the support of **SDAIA** can support retailers implement virtual try-on features for clothing and accessories, allowing customers to see how products look on them in real-time.
Education	**Enhance the educational experience** in Saudi Arabia by providing **interactive and personalized learning content**, making education more **engaging** and **accessible**.	The **Ministry of Education** can deploy virtual real tutors in remote areas, providing personalized and interactive lessons to students.
Healthcare	**Revolutionize** Saudi's **medical training** and **patient education** by providing realistic **simulations** and **voice reconstruction** possibilities	The **Ministry of Health** with **SDAIA** can implement deepfake technology in medical schools to provide students with realistic training simulations
Culture	**Preserve and revive Saudi cultural heritage** through lifelike **historical reenactments** and **preservation** of endangered dialects	The **Ministry of Culture** can collaborate with **SDAIA** to create deepfake reenactments of pivotal moments in Saudi history, such as the unification of the Kingdom or the life of King Abdulaziz Al Saud

Figure 5: *Non-malicious examples of deepfakes across sectors.*

Below, we explore sample examples on how these sectors can benefit from the innovative and valuable applications of deepfakes

9.1. Marketing

Personalized Ads: Deepfakes enable the creation of highly personalized advertisements featuring famous individuals, provided there is explicit consent. This approach can lead to increased engagement and conversion rates by resonating more deeply with target audiences.

Dynamic Content: Marketers can use this technology to generate realistic promotional content tailored to specific demographics at a lower cost. This allows for the rapid production of diverse marketing materials that can adapt to audience preferences.

Virtual Influencers: Virtual influencers, created using deepfake technology, can promote products and services effectively. These AI-generated personas can engage with audiences in innovative ways, providing consistent brand messaging without the constraints faced by human influencers.

Below is a sample list of Do’s and Don’ts to guide the ethical application of deepfakes in the marketing sector.

Do’s:

- Ensure that you have clear, documented consent from individuals whose likeness or voice is being used in marketing materials.
- Use deepfake technology to make the figure endorse products or services they have agreed to promote.

Don’ts:

- Do not use deepfake technology in ways that could be perceived as manipulative or unethical, such as altering real-world events or statements to fit a marketing narrative.
- Do not use an individual's likeness without their permission, as this can lead to significant privacy violations and legal issues.

9.2 Entertainment

Film Production: Deepfakes can be used to create or alter scenes in films without the need for expensive reshoots or special effects. This technology can save significant time and costs in film production while maintaining high-quality visuals.

Podcast Development: The entertainment industry can leverage deepfakes to generate low-cost podcasts featuring the voices of renowned storytellers or celebrities, thereby attracting more listeners and expanding their reach.

Fan Engagement: Celebrities can use these tools to engage with their fans through personalized messages or interactions, enhancing fan loyalty and creating memorable experiences.

Below is a sample list of do’s and don’ts to help navigate the responsible use of deepfakes in the entertainment sector.

Do’s:

- Use deepfakes to enrich storytelling in films and media, enabling the seamless integration of visual effects and actor portrayals.
- Use deepfakes to create positive and memorable fan interactions, such as personalized messages from celebrities or virtual experiences that strengthen fan loyalty provided the public figure’s consent.

Don’ts:

- Steer clear of using deepfakes to create content that could be considered harmful, even if it was originally not intended.

9.3. Retail

Virtual Try-Ons: Deepfake technology allows customers to virtually try on clothes, accessories, and makeup, significantly enhancing the online shopping experience. This interactive feature can reduce return rates and increase customer satisfaction by providing a realistic preview of products.

Product Pitches: Companies can utilize deepfakes to pitch products to investors remotely, creating compelling presentations that increase the likelihood of securing funding. This approach can simulate face-to-face interactions, making pitches more persuasive.

Below is a sample list of do’s and don’ts to steer the ethical use of deepfakes in the retail sector.

Do’s:

- Use deepfakes to create realistic virtual try-ons for clothing, accessories, and makeup, improving the online shopping experience and reducing return rates.
- Do utilize deepfakes in product pitches to investors to create engaging and persuasive presentations, especially when remote communication is necessary.

Don’ts:

- Do not use deepfakes to present products in a way that misrepresents their appearance, functionality, or quality, leading to customer dissatisfaction.
- Avoid using deepfake technology to create overly invasive personalized ads or experiences that could make customers feel uncomfortable or violated.

9.4. Education

Interactive Learning: Deepfakes can create engaging and interactive educational content featuring personalized virtual tutors. This can enhance learning experiences by making them more immersive and tailored to individual student needs.

Virtual Classrooms: Deepfake technology can enhance online education through realistic avatars of teachers and students, fostering a more interactive and engaging learning environment.

Remote Training: Virtual trainers powered by deepfakes can facilitate remote training sessions, reducing costs and logistical challenges associated with traditional training methods.

Below is a sample list of do’s and don’ts to ensure that deepfakes are used responsibly in the education sector.

Do’s:

- Use deepfakes to develop interactive and immersive educational materials, making learning more engaging and effective for students.
- Use deepfake-powered virtual trainers to make remote training and education more accessible to a broader audience, including those in remote or underserved areas.
- When using deepfakes in educational content, ensure that the information presented is accurate and culturally respectful, promoting a well-rounded understanding of the subject.

Don’ts:

- Do not use deepfakes to create educational content that could spread false or misleading information, as this can have serious consequences for learners.
- Avoid over-reliance on deepfake technology as a substitute for real teachers or educators, as personal interaction and mentorship are crucial for effective learning.

9.5. Healthcare

Medical Training: Deepfakes can generate realistic medical images of patients, aiding in the training of doctors on diagnosis and treatment techniques. This can improve the quality of medical education and patient care.

Voice Reconstruction: For patients who have lost their ability to speak due to disease or accidents, this technology can regenerate their real voices, significantly improving their quality of life.

Health Campaigns: Health organizations can develop realistic and engaging health awareness campaigns using deepfakes to effectively communicate important messages and drive public health initiatives.

Below is a sample list of do’s and don’ts to guide the safe and ethical use of deepfakes in the healthcare sector.

Do’s:

- Use deepfakes to create realistic medical simulations for training purposes, helping medical professionals to improve their skills and knowledge.
- Utilize deepfakes to create realistic and compelling health awareness campaigns that resonate with diverse audiences and drive public health initiatives.

Don’ts:

- Do not use deepfakes to generate medical content or advice that could mislead patients or healthcare providers, potentially causing harm.
- Avoid using deepfakes to simulate patient outcomes or disease progression in a way that could give patients or families false hope or unwarranted fear.

9.6. Culture

Historical Reenactments: Deepfakes can be used to create lifelike reenactments of historical events, preserving cultural heritage in an engaging and educational manner. This can bring history to life and make it accessible to broader audiences.

Language and Dialect Preservation: This technology can help preserve and revive endangered languages and dialects by creating realistic voices. This application supports cultural preservation and linguistic diversity.

Below is a sample list of do’s and don’ts to promote the responsible use of deepfakes in cultural preservation.

Do’s:

- Use deepfakes to recreate and preserve cultural events and traditions, making them accessible to broader audiences and future generations.

- Apply deepfakes to revive and teach endangered languages and dialects, contributing to the preservation of cultural diversity and heritage.

Don’ts:

- Do not use deepfakes to inaccurately portray cultural practices, figures, or events. Misrepresentation can lead to cultural appropriation, misunderstanding, and disrespect.

- Do not exploit cultural heritage for commercial purposes without proper permission and respect for the source community. Ensure that any use of deepfakes in this context benefits the cultural community involved.

Harnessing the power of deepfakes for positive and ethical purposes can lead to remarkable advancements and deliver enhanced experiences across various sectors. By focusing on continuous learning and skill development, ensuring organizational preparedness, and emphasizing ethical and positive applications, the transformative potential of deepfake technology can be fully realized while minimizing associated risks.

The three main key takeaways are the following:

- Continuous Learning and Skill Development: Ongoing education and hands-on training are crucial for staying updated with AI developments, methodologies, and ethical considerations, ensuring effective management of AI applications.

- Organizational Preparedness: Developing internal AI skills through tailored workshops and strategic hiring practices enhances innovation and problem-solving capabilities, enabling organizations to effectively handle AI and deepfake technologies.

Ethical and Positive Applications: Focusing on non-malicious uses in sectors such as marketing, entertainment, retail, education, healthcare, and culture can drive significant advancements and deliver enhanced experiences while responsibly managing risks.

10. Conclusion

Deepfake technology presents both significant risks and substantial opportunities across sectors. The ethical and responsible use of deepfakes can lead to notable advancements in marketing, entertainment, retail, education, healthcare, and cultural applications, enhancing user experiences and driving innovation. The guidelines provided emphasize continuous learning and skill development, urging organizations to develop internal AI capabilities through tailored workshops and strategic hiring practices to effectively manage deepfake technologies.

Ethical considerations are paramount in deploying deepfake technologies. Developers and content creators must prioritize privacy, transparency, accountability, and social benefits to mitigate risks. Implementing strong data protection measures, securing consent, and maintaining transparency can build trust and prevent misuse. Deepfake technology developers should also ensure their tools include mechanisms to trace and verify content authenticity to prevent misuse. Consumers should adopt robust verification processes and stay informed about the characteristics and detection of deepfakes to protect themselves and others.

Public awareness and education are critical in combating potential harms from deepfakes. Through awareness campaigns and educational initiatives, individuals can better recognize and respond to deepfakes, reducing the risk of manipulation and misinformation. By adhering to these recommendations, stakeholders can ensure deepfake technology is used beneficially and ethically, safeguarding public trust and promoting responsible innovation. Through collaborative efforts and a commitment to ethical practices, the transformative potential of these tools can be harnessed to drive positive change across multiple domains.

11. Appendix

11.1. Appendix 1: Definitions

Deepfakes: Hyper-realistic synthetic media created using deep learning techniques, to manipulate audio, video, or other digital content to convincingly alter the likeness, actions, or appearance of beings, objects, or the depiction of events, making it difficult to distinguish between real and fake content.

Malicious Deepfakes: Deepfakes designed to deceive, harm, or exploit individuals, often used for spreading false information, defamation, or fraud.

Non-malicious Deepfakes: Deepfakes utilized for beneficial purposes, without any intent to deceive or harm. Examples include marketing, entertainment, retail, education, healthcare, and cultural applications.

Deepfake Technology Developer: An individual or entity that creates the technology used for generating deepfakes, focusing on the development of AI algorithms and tools that enable the creation of synthetic media.

Deepfake Content Creator: An individual or entity that employs deepfake technology to create digital content such as images, videos, or audio for various purposes, including content creators, businesses, educators, and researchers.

Deepfake Consumer: Any individual or organization that interacts with or is exposed to deepfakes, including citizens, employees, companies, and organizations.

11.2. Appendix 2: Training and Skills Development

In the rapidly evolving landscape of AI and deepfake technology, continuous learning is essential to stay abreast of new developments, methodologies, and ethical considerations. Continuous learning enables individuals and organizations to adapt to technological advancements, improve their skill sets, and remain competitive.

For end-users, ongoing education helps ensure that they can effectively develop, implement, and manage AI applications.

Workshops are crucial for equipping end-users with the skills and knowledge needed to combat malicious uses of this technology.

These workshops should cover various aspects, including:

1. **Understanding Deepfake Technology:** Providing end-users with a comprehensive understanding of how deepfakes are created and the underlying AI algorithms.

2. **Detection Techniques:** Teaching advanced techniques for identifying deepfakes, such as analyzing inconsistencies in video and audio, using AI-based detection tools, and understanding the telltale signs of synthetic media.

3. **Ethical and Legal Considerations:** Discussing the ethical implications and legal frameworks surrounding the use of deepfakes to ensure responsible and compliant practices.

4. **Hands-on Training:** Offering practical sessions where end-users can work with detection tools and datasets to gain hands-on experience in identifying and mitigating deepfake threats.

To ensure that end-users remain up-to-date with the latest advancements and best practices, it is recommended that these trainings / workshops are conducted at regular intervals. A proposed schedule might include:

- Semi-annual Workshops: Conducting workshops every six months to provide updates on the latest developments in deepfake technology and detection techniques.

- Annual Refresher Courses: Offering comprehensive annual refresher courses to reinforce key concepts, review new ethical and legal considerations, and ensure continuous competency in using detection tools.

- Special Sessions: Organizing special sessions following significant technological breakthroughs or changes in regulatory frameworks to keep participants informed and prepared to adapt to new challenges.

By adhering to this training schedule, end-users can maintain a high level of preparedness and effectiveness in addressing the threats posed by deepfake technology. Moreover, some targeted The training guidelines can be followed by government entities, developers, and the general public to ensure effective education and resilience against the misuse of deepfake technology.

11.2.1. Government Entities:

The purpose of these guidelines is to equip government officials with the knowledge and tools needed to understand, detect, and respond to deepfake threats, ensuring the integrity of public communications and security.

Training should focus on deepfake technology, detection techniques, legal frameworks, policy implications, and case studies. Officials need a comprehensive understanding of how deepfakes are created, the ethical and legal considerations surrounding them, and the potential impact on national security and public trust.

It is recommended that government entities participate in quarterly workshops to stay updated on the latest developments. Additionally, bi-annual hands-on training sessions with advanced detection tools should be conducted to provide practical experience and enhance their ability to respond effectively to deepfake threats.

11.2.2. Developers:

This section aims to guide developers in building a robust understanding of deepfake technology, ensuring they can create and detect deepfakes responsibly, and integrate effective detection tools into their software solutions.

Training should emphasize the technical aspects of creating and detecting deepfakes, ethical AI development practices, and the integration of detection tools in software development. Advanced sessions should cover AI algorithms, machine learning techniques, and hands-on coding exercises.

Developers should engage in monthly technical webinars to stay informed about the latest advancements in AI and deepfake technology. Bi-annual intensive workshops should include hands-on coding sessions, advanced AI algorithms, and problem-solving exercises to ensure developers are well-equipped to tackle deepfake challenges.

11.2.3. General Public & Organizations:

The goal of these guidelines is to raise awareness among the general public and the private sector organizations about the risks of deepfakes and provide them with the knowledge and tools to identify and protect themselves from malicious content.

Public education should focus on raising awareness about the existence and dangers of deepfakes. Training should teach basic detection techniques, the importance of verifying information, and ways to protect personal data online. Practical advice on recognizing and reporting suspected deepfakes is essential.

Bi-annual public awareness campaigns should be conducted to highlight the risks and signs of deepfakes. Monthly informational webinars can provide basic detection techniques and advice on protecting personal data. Quarterly community workshops should offer interactive sessions on recognizing and responding to deepfakes.

By tailoring the training content and frequency to the specific needs of government entities, developers, and the general public, stakeholders can effectively address the challenges posed by deepfake technology. Continuous education and proactive engagement are key to maintaining resilience against the misuse of AI and deepfake technologies.

11.3. Appendix 3: Case Studies and Examples

11.3.1 Deepfake Harmful Scenarios:

Imposter scam example[12]: Kidnapping scam in South Asia

- In a South Asian country, a scammer used advanced AI voice cloning technology to impersonate a child’s voice. The scammer called the child’s parents, claiming that their child had been kidnapped and demanded a ransom. In a moment of panic, the parents transferred a significant amount of money via a digital payment platform.

- The scam caused severe emotional distress to the parents, financial loss to the family, and heightened public awareness of the dangers of deepfake technology. The scammer exploited advanced technology unethically to manipulate and deceive individuals, causing harm and distress. This incident underscores the critical importance of digital literacy and the necessity for individuals to verify unexpected communications as a defense against such sophisticated scams.

- The local police launched an investigation to trace the digital payment and identify the caller. Authorities issued public warnings about AI-driven scams, advising citizens to use family-specific code words for verification. They also collaborated with tech companies to understand the scam's tools and develop prevention strategies.

Non-consensual manipulation example[13]: Unauthorized deepfake scandal

- In 2023, a prominent political figure in the United States spoke out about her encounter with a synthetic image of herself on a social platform. The deepfake image, created without her consent, caused shock and resurfaced past traumas. She stated that deepfakes "parallel the same exact intention of violent humiliation."

- The incident not only caused significant psychological trauma to the victim but also eroded public trust, prompted potential legal actions against the creators, and increased both public awareness and legislative interest in regulating deepfake technologies. Creating and distributing a non-consensual deepfake image of a public figure undermines ethical norms by disrespecting individual autonomy and privacy, and it inflicts emotional harm. This case highlights the urgent need for ethical guidelines and stringent legal frameworks to govern the creation and dissemination of synthetic media to protect individuals from harm and preserve societal trust.

- The incident spurred discussions about the need for stronger regulations against the misuse of deepfake technology. Efforts included drafting new laws aimed at preventing the creation and distribution of harmful deepfake content. Federal and state legislators proposed various bills targeting deepfakes, particularly focusing on their use in political campaigns and protecting individuals from impersonation and fraud. These legislative efforts were part of a broader push to address the evolving threat of AI-driven misinformation and protect public figures and citizens alike.

[12] Times of India
[13] The Guardian

Disinformation and propaganda example[14]: Fake audio clip scandal

- A faked audio clip with the voice of a political leader from a liberal party in a European country circulated on social media. The clip falsely suggested that the politician had been buying votes from a minority group in the country.
- The fake audio clip not only manipulated public opinion and potentially influenced important decisions but also damaged the reputation of the individuals involved and prompted calls for regulatory reforms and better detection tools for deepfakes. The creation and dissemination of the faked audio clip breached ethical norms by spreading misinformation, deceiving the public, and undermining trust. This incident underscores the crucial need for heightened digital literacy and robust verification tools to safeguard against the misuse of synthetic media, thereby protecting societal trust and integrity.
- The incident prompted calls for stronger regulations and policies to combat deepfakes and other synthetic media. Legislative bodies in the country began discussing new laws to prevent the creation and spread of such deceptive content. At the time of the incident, the country did not have specific laws addressing deepfakes, so existing defamation and cybercrime laws were invoked. Following the incident, there was increased momentum to draft and implement comprehensive legislation specifically targeting deepfakes and synthetic media to protect public figures and the democratic process from similar attacks in the future

11.3.2 Deepfake Beneficial Scenarios:

Marketing application[15]: Deepfake ad campaign

- A popular food delivery platform harnessed deepfake technology to create a captivating and context-aware advertisement. The ad features a famous actor, who appears to be ordering food based on his cravings. However, the ad changes based on the viewer’s GPS location. The seamless integration of the actor’s likeness and the personalized content made this ad engage viewers on a local and relatable level.
- The use of deepfake technology in the ad campaign resulted in cost-effective personalization, increased viewer engagement, and higher conversion rates, significantly boosting the platform's revenue and brand loyalty through innovative advertising strategies. The campaign exemplifies an ethical application of deepfake technology in advertising, characterized by consent, transparency, and disclosure. This example demonstrates the potential of deepfake technology as a powerful tool for personalized and dynamic advertising, highlighting the importance of maintaining ethical standards such as consent, disclosure, and transparency to foster trust and authenticity in digital marketing.

Entertainment application[16]: De-aging in film

- A film directed by a renowned director features a well-known actor in the lead role as a hitman. The movie required flashback scenes where the actors needed to appear younger. To achieve this effect, the filmmakers used deepfake technology to de-age the lead actor and others, digitally altering their appearances and giving them a more youthful look.
- The use of deepfake technology in the film resulted in significant cost savings and production efficiencies, allowed for high-quality realistic de-aging effects, and generated substantial media buzz, likely boosting box office performance and audience engagement.

[14] The bureau of Investigative Journalism
[15] Analytics Drift
[16] Digital Trends

- This application of deepfake technology was entirely consensual and approved, showcasing an ethical and beneficial use of the technology in artistic expression to enhance storytelling and viewer experience. The film demonstrates how deepfake technology can be harnessed creatively and responsibly in the entertainment industry, highlighting the potential for technological innovations to enhance artistic endeavors while maintaining ethical standards.

Healthcare application[17]: ALS voice reconstruction.

- Voice reconstruction using deepfake technology has shown promising results for patients with Amyotrophic Lateral Sclerosis (ALS), a neurodegenerative disease that affects nerve cells in the brain and spinal cord, causing patients to lose their ability to move and speak.
- Voice reconstruction using deepfake technology significantly improved the quality of life for patients with ALS, enabling them to regain their ability to communicate, continue their careers, and maintain their personal relationships, while also inspiring new projects to support others with similar conditions. This application of deepfake technology exemplifies a beneficial and ethical use, enhancing the lives of patients with ALS by restoring their ability to communicate and maintaining their quality of life. The use of deepfake technology in medical applications can provide significant benefits, demonstrating its potential to improve the quality of life for patients with severe communication impairments while maintaining ethical standards.

11.4. Appendix 4: Risk Assessment Framework for Deepfake Technology Developers

11.4.1 Phase 1: Plan & Design

Phase	Question	Principle	Risk Mitigation
PD.1	Did you design the appropriate level of human oversight for the AI system and use case?	Accountability & Responsibility	Ensure that human oversight is embedded at all critical stages to prevent misuse
PD.2	Does your AI system design prevent overconfidence in the AI system with necessary human intervention mechanisms?	Accountability & Responsibility	Implement fail-safes and human intervention points to correct any AI misjudgments
PD.3	Did you define human oversight processes with the appropriate KPIs and assign responsibility to relevant parties?	Accountability & Responsibility	Regularly review and update KPIs, ensuring clear accountability
PD.4	Did you design an operation and governance strategy to abort or intervene in the system when it doesn't work in an intended way?	Accountability & Responsibility	Develop and test response protocols for system failures or deviations
PD.5	Did you consider liability and Data Subject protection requirements and take them into account?	Privacy & Security	Comply with data protection laws and secure necessary consents
PD.6	Did you define thresholds of the KPIs and did you put governance procedures or autonomous actions in place to trigger alternative/rollback plans?	Accountability & Responsibility	Establish and monitor threshold indicators to activate contingency plans
PD.7	Did you provide training and education to help develop accountability practices?	Accountability & Responsibility	Conduct regular training sessions on ethical standards and accountability

[17] CBS News

11.4.2. Phase 2: Prepare Input Data

Phase	Question	Principle	Risk Mitigation
PID.1	Is there an established mechanism that flags issues related to data privacy or protection in the process of data collection and processing	Privacy & Security	Implement automated and manual checks for data privacy compliance
PID.2	Has the data been reviewed in terms of scope and categorization?	Privacy & Security	Categorize data appropriately and restrict access to sensitive data
PID.3	Has the data been reviewed to check if personal data is evident within the dataset? Is there an established mechanism that allows the AI model to train without use of personal or sensitive data?	Privacy & Security	Use anonymization and pseudonymization techniques where applicable

11.4.3. Phase 3: Build & Validate

Phase	Question	Principle	Risk Mitigation
BV.1	Has the behavior of the system been tested against unexpected situations and environments? Is there a defined fallback plan if the AI model encounters adversarial attacks or other unexpected situations?	Reliability & Safety	Conduct thorough testing under varied scenarios and establish robust fallback mechanisms
BV.2	Are there defined processes that outline procedures to describe actions to be taken when an AI system fails in different contexts?	Reliability & Safety	Develop clear protocols and test them regularly
BV.3	Is there an established mechanism of communication to assure the end-users of the system's reliability?	Transparency & Explainability	Maintain open communication channels and provide regular updates on system performance
BV.4	Are there clear and understandable definitions explaining the outcomes of the Deepfake system?	Transparency & Explainability	Ensure that the deepfake process is documented and understandable to users

11.4.2. Phase 2: Prepare Input Data

Phase	Question	Principle	Risk Mitigation
DM.1	Has the team assessed the AI system's vulnerabilities to potential attacks, revelation of sensitive data, or breaking the confidentiality?	Privacy & Security	Perform regular security audits and vulnerability assessments
DM.2	Are there mechanisms to measure if the system is producing an unacceptable amount of inaccurate results?	Accountability & Responsibility	Implement continuous monitoring and quality control measures
DM.3	Are the persons who are accessing the data qualified with the necessary competences to understand the details of data protection requirements?	Privacy & Security	Ensure personnel have appropriate training and certification in data protection

11.5. Appendix 5: Sample Consent Form for the Use of Personal Data in Deepfake Content Creation

11.5.1. Purpose of the Consent

This consent form outlines how your personal data, including images, video, and/or audio, will be used to create synthetic media content (deepfakes). The content generated will be used for [specific purpose, e.g., educational videos, public awareness campaigns, research, etc.]. Please review this document carefully before providing your consent.

11.5.2. Data to be Used

[List specific data types: e.g., Photographs, Video footage, Audio recordings, etc.]

11.5.3. Scope and Duration of Use

The data provided will be used solely for the purpose of [specific purpose].

The content created will be used from [start date] to [end date or "ongoing until revoked"].

The data will be stored securely and only accessible by authorized personnel.

11.5.4. Right to Withdraw Consent

You have the right to withdraw your consent at any time. To withdraw consent, please contact [contact information]. Upon withdrawal, your data will no longer be used in future deepfake content creation.

11.5.5. Potential Risks

While all reasonable measures will be taken to protect your data, there may be inherent risks in the use of personal data for synthetic media. These risks include [outline potential risks, e.g., misuse by unauthorized parties, potential for deepfake content to be misinterpreted, etc.].

11.5.6. Confirmation of Consent

By signing below, you confirm that you have read and understood the information provided in this consent form. You agree to the use of your personal data as outlined above.

[Checkbox] I consent to the use of my personal data for the creation of deepfake content as described.

[Checkbox] I acknowledge that I have been informed of my right to withdraw consent at any time.

Name of Data Subject: ___________________________

Signature of Data Subject: _______________________

Date: ____

Contact Information for Withdrawal or Inquiries:

- [Name of Contact Person]
- [Email Address]
- [Phone Number]

SDAIA

الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority