CohesionX Conceptual Framework for Ethical Issues in Generative AI

Client(s): Internal

Product(s): Generative AI

Document Control

The document is under version control at CohesionX. Any alterations will be under version control and must be authorised by the AI Committee

Document information	
Document description	This document describes the Cohesion X Ethical Framework for AI, addressing the ethical concerns associated with businesses and corporations utilising Generative AI.
Filename	CohesionX Conceptual Framework for Ethical AI.docx

© CohesionX (PTY) LTD

Table of Contents

© CohesionX (PTY) LTD

1. ABSTRACT

This document comprehensively explores Artificial Intelligence (AI), specifically focusing on Generative AI and Large Language Models (LLMs). It examines their transformative role across various sectors, including healthcare, agriculture, civil rights, human resources management, and insurance. The introduction sets the stage by discussing AI's significance in the fourth industrial revolution and the ethical, legal, and economic challenges it presents. The document further delves into the structure of Generative AI, outlining essential elements such as generative models, training data, and various technical components crucial for these systems' functioning.

Significant emphasis is placed on the applications and use cases of Generative AI and LLMs, highlighting their profound impacts and limitations. Examples like ChatGPT demonstrate these technologies' capabilities in understanding relationships within large datasets while revealing their limitations in discerning truth and reliability. The text also addresses the ethical considerations inherent in Generative AI technology, touching upon issues like authenticity, bias, transparency, accountability, intellectual property, economic impact, and the erosion of human skills.

A notable section of the document is dedicated to the European Union's human-centric approach to AI, contrasting it with strategies from the US and China. This approach emphasises respect for human values throughout the lifecycle of AI systems. Additionally, the text introduces the Cohesion X Framework, a set of ethical standards designed to guide the responsible development and deployment of AI technologies. This framework includes principles like human agency and oversight, technical robustness, privacy, transparency, diversity, societal well-being, and accountability.

In conclusion, the document underscores the necessity of a balanced, ethical, and cautious approach to deploying and using Generative AI and LLMs. It calls for continued vigilance and responsible management to harness these technologies' benefits while mitigating risks and limitations.

2. INTRODUCTION

Artificial Intelligence (AI) broadly encapsulates methods in machine learning for extensive data analysis, robotics focused on the development and functionality of programmable machines, and algorithms and automated decision-making systems (ADMS) with proficiency in predicting actions of humans and machines, subsequently making autonomous decisions. AI technologies are pivotal for economic and social growth, finding applications in diverse sectors like healthcare for breakthrough cancer treatment discovery and transportation for traffic pattern prediction and aiding autonomous vehicles to optimise energy and water consumption. The pervasiveness of AI in our daily lives is escalating, solidifying its status as a central force in what's often termed the fourth industrial revolution.

Despite the consensus on AI's myriad benefits, addressing its ethical, legal, and economic dilemmas is imperative, particularly concerning potential violations of human rights and fundamental freedoms. AI poses significant risks, including the security of personal data and privacy and the hazard of discrimination through profiling individuals or managing specific scenarios via algorithms.

In the criminal justice system, AI's impact is a subject of intense debate, focusing on the repercussions of AI and robotic technologies on employment, including job losses due to increased automation. Additionally, the need to scrutinise the role of algorithms and ADMS in contexts such as defective products (safety and liability), digital finance (blockchain technology), the spread of

© CohesionX (PTY) LTD

misinformation (fake news), and potential military applications (autonomous weaponry and cybersecurity) is gaining attention.

Moreover, the challenge of instilling ethical principles in developing and designing algorithms and AI systems is ongoing.

3. STRUCTURE

3.1. Generative AI

Generative Artificial Intelligence (AI) refers to a subset of AI technologies capable of creating new content, data, or outputs extending beyond their original training data.

These advanced AI systems are adept at producing outputs that mirror human-created works across various fields, including art, music, medicine, and journalism. They have the capability to craft unique art pieces, assist in design, compose music, simulate novel medical treatments, forecast disease patterns, aid in journalistic reporting, and perform data collation.

3.2. The essential elements of Generative AI encompass:

1. **Generative models:** These algorithmic frameworks are developed to produce data that resembles their training datasets. Generative Adversarial Networks (GANs) are a predominant example in this category.

2. **Training data:** This refers to the large and diverse datasets that are crucial for training generative models. The quality and variety of this data significantly influence the AI's performance.

3. **Generative Adversarial Networks (GANs):** Comprising two neural networks—the Generator and the Discriminator—GANs work in tandem. The Generator aims to create data, while the Discriminator strives to differentiate between real and generated data.

4. **Recurrent Neural Networks (RNNs):** These are mainly used in sequential data generation tasks like text generation, where the order of input is crucial.

5. **Neural Network Architecture:** This defines the structure and layout of the deep learning model, influencing how the AI processes information and learns from data.

6. **Variational Autoencoders (VAEs):** These probabilistic models learn from a compressed representation of training data and can generate new examples by sampling from this learned space.

7. **Loss functions:** These mathematical functions measure the difference between the model's predictions and actual outcomes, aiding in parameter adjustment during training.

© CohesionX (PTY) LTD

8. **Latent Space:** A reduced representation of input data, typically in a lower-dimensional space, from which generative models can create new instances.

9. **Transfer Learning:** This approach uses pre-trained models for new, related tasks, leading to faster and more efficient training of generative models.

10. **Optimisation algorithms:** These are used to refine neural network parameters, guided by feedback from the loss functions to enhance model performance.

11. **Regularisation:** A method to prevent overfitting, ensuring the model generalises well and doesn't simply memorise the training data.

12. **Synthetic Data Generation:** This involves using generative models to create artificial data when actual data is scarce or unavailable.

13. **Evaluation metrics:** Tools used to assess the quality and diversity of the generated outcomes.

14. **Feedback loops:** A process where outputs from a generative model are fed back as inputs, allowing for iterative improvement or continuous production.

3.3. Application and Use Cases of Generative AI and LLM

In 2017, Google unveiled a transformative technology known as Transformers, revolutionising how machines process sequential data like text, audio, or video. These Transformers laid the groundwork for Large Language Models (LLMs), some containing trillions of parameters. Initially designed for text conversion, LLMs evolved to generate diverse data formats, including images and videos. Trained on vast datasets, they can perform varied tasks, leading to their designation as 'Foundation Models', crucial in the generative AI field.

An intricate training process lies at the core of generative AI applications like ChatGPT. ChatGPT's development involved analysing text data (source undisclosed by OpenAI) to understand word relationships. Its training involved three phases: initial supervised learning with human review, followed by a phase where human rankings of model responses refined the model, and a final large-scale reinforcement learning round. This method exemplifies Foundation models' ability to learn tasks beyond their direct training.

However, LLMs like ChatGPT, despite appearing intelligent, only reflect statistical knowledge without proper understanding. They're sometimes referred to as 'stochastic parrots', a term from a controversial 2021 paper, due to their lack of real comprehension. They excel by processing large-scale data and parameters, but this also leads to limitations. LLMs can propagate misinformation and biased content as they can't discern truth or appropriateness in their outputs. Their probabilistic nature means varied responses to the same prompt, posing reliability issues.

Large Language Models (LLMs) are emerging as transformative tools across diverse sectors, each demonstrating unique applications and benefits despite their current limitations. From healthcare to agriculture, civil rights to human resources management, and the short-term insurance industry, LLMs are making significant strides, indicating a future rich with innovation and efficiency.

© CohesionX (PTY) LTD

In the healthcare sector, LLMs are revolutionising practices with their ability to generate medical reports, significantly reducing documentation time and aiding in medical education. Their impact extends to drug discovery and biomolecule design, as exemplified by Insilico Medicine's pioneering work. Major tech firms, such as Microsoft, recognise this potential and actively invest in LLM integration, even in the face of current challenges.

The agriculture sector also benefits significantly from LLMs. These models enhance crop data analysis, optimise farming techniques, and streamline supply chain management. They are crucial in precision agriculture, employing data analytics for tasks like soil health assessment and crop monitoring. Their role in predictive analytics is particularly invaluable, aiding in forecasting weather patterns and pest infestations. Agritech companies are rapidly adopting LLMs to boost productivity and sustainability, with major tech firms following suit and investing despite acknowledged limitations.

LLMs offer invaluable tools for civil rights organisations as well. They efficiently analyse large volumes of legal documents and case law, supporting the development of legal strategies for civil rights litigation. These models are adept at monitoring social media and news reports to detect civil rights violations, enhancing prompt response capabilities. In drafting and reviewing policy proposals, LLMs ensure alignment with civil rights objectives. They also play a crucial role in public education and advocacy, producing clear, accessible information on civil rights issues. This automation allows organisations to focus more resources on direct action and advocacy, although careful usage is advised, given the current limitations of the technology.

In corporate human resources management, LLMs are proving to be indispensable. They streamline HR processes by automating tasks like resume screening and candidate assessments, reducing bias and saving time. Personalised training and development programs, tailored through extensive data analysis, enhance employee engagement. LLMs are also crucial in analysing employee feedback and fostering improved workplace culture and satisfaction. In talent acquisition and retention, they provide advanced analytics for better decision-making. Recognising these advantages, many large corporations are beginning to integrate LLMs into their HR operations despite existing technological constraints.

In the short-term insurance industry, LLMs are emerging as game-changers. They improve claim processing efficiency by quickly analysing and validating claims, enhancing customer satisfaction. Personalised policy customisation, enabled through data analytics and advanced fraud detection capabilities, exemplifies their impact. Additionally, LLMs offer round-the-clock customer support, handling inquiries and guiding customers on policy and claims procedures. They also contribute significantly to risk assessment, analysing extensive datasets for more accurate insurance product pricing. The integration of LLMs in this industry is seen as a crucial step towards more efficient, customer-focused services, even with the current limitations.

Across these sectors, LLMs demonstrate a remarkable ability to enhance efficiency, accuracy, and innovation. As they continue to evolve, these models promise even more significant impacts, ushering in an era of heightened productivity and advanced data-driven decision-making across various industries.

When implemented across a wide range of sectors, generative AI systems and large language models (LLMs) demand careful and thoughtful application. Examples such as Meta's Galactica bot have highlighted the potential for these technologies to produce misleading or inaccurate information. Integrating ethical considerations and ensuring persistent human supervision in their utilisation is imperative. The risks associated with misuse and excessive dependence on these advanced technologies are notable. There have been numerous incidents where reliance on AI and LLMs has led to the widespread dissemination of misinformation, underlining the need for vigilance and responsible management in their deployment.

© CohesionX (PTY) LTD

LLMs should be approached as evolving tools needing stringent supervision. They can significantly improve efficiency but require careful handling. As generative AI grows, a balanced, ethical, and cautious approach is essential in healthcare applications, ensuring safety and utility while mitigating risks and limitations.

4. ETHICAL CONSIDERATIONS IN GENERATIVE AI TECHNOLOGY

In the realm of Generative AI technology, ethical considerations encompass a broad range of societal, legal, and moral issues pertinent to the application of AI systems designed for content creation.

The primary ethical dimensions of Generative AI include:

- **Authenticity and Veracity:** This area focuses on the challenges presented by the blurring line between genuine human-generated content and AI-produced materials, such as deepfakes. Such AI creations can be deceptively realistic, posing risks to truth in areas like journalism, politics, and public perception. The autonomy of Generative AI in creating content raises questions about the originality and potential for unintentional plagiarism, underscoring the necessity for tools to verify content authenticity.

- **Bias and Fairness in Data:** Ensuring that the datasets training Generative AI are free from biases is critical to avoid discriminatory outcomes. The potential for AI systems to perpetuate or amplify existing societal biases necessitates measures to ensure impartiality, promoting fairness and equity in AI applications.

- **Transparency and Explainability:** These principles are vital in clarifying the AI decision-making process, especially when AI outcomes have significant real-world impacts. They involve disclosing information about the AI model's functioning and providing understandable rationales for its outputs, ensuring that explanations align with the model's actual decision-making process.

- **Accountability and Responsibility:** This aspect involves identifying who is liable when a Generative AI system produces harmful or misleading content. It includes ethical and legal considerations, emphasising the need for clear standards and processes to manage the repercussions of AI-driven outputs.

- **Intellectual Property and Artist Rights:** Concerns here revolve around the ownership of AI-generated content and the rights of creators in content development. Issues like unintentional IP infringement and the economic distribution of benefits from AI-generated works are also critical.

- **Economic and Social Impact:** This encompasses the potential job creation and displacement resulting from Generative AI and its influence on the valuation of human-created content. The impact on education and the risk of spreading misinformation through deepfakes are also significant considerations.

- **Privacy and Security Considerations:** These focus on safeguarding personal information in the context of AI-generated content that can replicate or exploit such data. Addressing these issues is crucial to ensure AI's safe and respectful usage.

© CohesionX (PTY) LTD

- **Erosion of Human Skills and Reliance on AI:** Generative AI's capacity to autonomously perform tasks traditionally requiring human labour raises concerns about the diminishing opportunities for skill development and an over-reliance on automation, which could lead to the devaluation of human intuition and traditional skills.

5. FRAMEWORK

5.1. Embracing a Human-Centric Perspective in Artificial Intelligence

Our guidelines are grounded in the European Union's human-centric approach to AI. This approach aligns with European values and principles, emphasising respect and adherence to these ideals.

Central to the EU's stance is the prioritisation of human values in the lifecycle of AI systems - from their development and deployment to their usage and oversight. This ensures compliance with the fundamental rights established in the EU Treaties and the Charter of Fundamental Rights of the European Union. These rights are founded on the principle of human dignity, acknowledging the unique moral status of humans. Additionally, the approach encompasses considerations for the natural environment and other living beings, advocating for sustainability and the welfare of future generations.

In the global AI race, the EU's strategy distinctly contrasts with the approaches of the US and China. Private sector innovations and self-regulation predominantly power the US model, whereas China's strategy is heavily government-directed, integrating both private and public investments in AI. The EU's approach is characterised by its commitment to cultural values and heightened safeguarding against AI-related societal risks, particularly in privacy, data protection, and anti-discrimination measures. This commitment distinguishes the EU from other nations with more relaxed regulations in these domains.

The EU's ethical guidelines for AI are designed to foster trustworthiness in AI systems. This entails ensuring compliance with all applicable laws and ethical standards and achieving both technical and social robustness to prevent unintentional harm. Furthermore, the guidelines underline the importance of centring AI systems around human needs. This means these systems should be developed, executed, and utilised in alignment with the ethical principles highlighted above.

5.2. Ethical Principles guiding our framework

To advance artificial intelligence (AI) responsibly and ethically, we align our framework with seven core ethical principles. These principles are in harmony with the comprehensive ethical considerations and frameworks established by the European Union (EU) for AI. The EU's approach, deeply rooted in respect for human dignity, privacy, fairness, and accountability, serves as the cornerstone for our guidelines. Our framework encapsulates these pivotal principles to ensure that AI technologies' deployment, development, and utilisation adhere to legal mandates and uphold the highest ethical values. The following key requirements reflect our commitment to fostering AI systems that are trustworthy, equitable, and beneficial to society as a whole, seamlessly integrating into the fabric of our ethical AI aspirations.

1. **Human Agency and Oversight:** The EU's ethical framework places great importance on upholding human autonomy and fundamental rights. In order to guarantee this in practical

© CohesionX (PTY) LTD

terms, it is advisable to conduct a fundamental rights impact assessment prior to the creation of artificial intelligence. Furthermore, AI systems should be specifically developed to facilitate human control and enable users to comprehend and engage with these systems efficiently. It is crucial to maintain the ability for humans to have supervision and the power to override AI decisions to prioritise human control.

2. **Technical Robustness and Safety:** Trustworthy AI requires that algorithms exhibit security, dependability, and resilience throughout all life cycle stages. This encompasses the tasks of dealing with cybersecurity, reducing the chances of cyber-attacks, and guaranteeing human supervision during emergencies. The European Union promotes cooperation between the artificial intelligence and security sectors and is contemplating legislation changes to tackle the responsibilities linked to AI.

3. **Privacy and Data Protection:** Adherence to the General Data Protection Regulation (GDPR) is vital, and artificial intelligence (AI) systems must be engineered to safeguard privacy and personal data. This involves implementing data encryption and anonymisation methods and assuring data accuracy by minimising biases and mistakes. Data-gathering procedures must adhere to impartiality and be subject to rigorous supervision.

4. **Transparency:** Maintaining transparency is crucial in order to prevent biases in AI. Key features include documenting datasets and procedures, identifying AI systems, and ensuring user knowledge of their interaction with AI. Explainability is crucial in ensuring that AI systems and the actions made by humans in relation to them can be comprehended and traceable.

5. **Promotion of Diversity, Non-Discrimination, and Fairness:** AI systems must be designed to prevent unfair prejudices and consider a wide range of human abilities, skills, and needs. This entails engaging and including stakeholders who could be directly or indirectly impacted by AI systems while also guaranteeing accessibility and impartiality.

6. **Societal and Environmental Well-being:** AI should actively foster positive social transformation and uphold environmental stewardship. Efforts to mitigate the environmental consequences of AI systems and evaluate their societal implications, such as their impact on physical and mental well-being, are strongly recommended. It is important to also take into account the ramifications of artificial intelligence on society and democracy, especially its impact on electoral processes.

7. **Accountability:** It is important to establish explicit methods to ensure that the responsible parties are held accountable for the effects of AI systems. This encompasses the examination of the ethical and legal consequences of AI, guaranteeing openness in decision-making procedures, and creating guidelines for accountability in cases of detrimental AI conduct.

5.3. Ethical Standards for implementing our framework

The comprehensive Cohesion X Framework has been meticulously crafted to address the multifaceted ethical quandaries inherent in the rapidly evolving domain of generative artificial intelligence. This framework, distinguished by its thoroughness and foresight, encompasses a range of pivotal domains, each underpinned by foundational standards that guide the responsible development and deployment of generative AI technologies.

© CohesionX (PTY) LTD

a. Articulation of Purpose and Contextual Precision

b. Transparency and Accountability Through Comprehensive Documentation

c. Data Rights and Privacy

d. Implementation of Authenticity Protocols

e. Ensuring Security and Robustness

f. Conducting Fairness Audits

g. Economic and Social Impact Assessment

h. Promotion of Public Engagement and Collaborative Efforts

i. Adaptability and Continuous Evolution

© CohesionX (PTY) LTD

Cohesion X GenAi Standards	Articulation of Purpose and Contextual Precision	Transparency and Accountability Through Comprehensive Documentation	Data Rights and Privacy	Implementation of Authenticity Protocols	Ensuring Security and Robustness	Conducting Fairness Audits	Economic and Social Impact Assessment	Promotion of Public Engagement and Collaborative Efforts
Generic definition for GenAI projects and products across most vertical implementations	Paramount to the framework is the establishment of a lucid and precise articulation of the intended purpose behind each generative AI system. This encompasses a deep understanding of such systems' varied contexts and applications, ensuring alignment with the established societal norms and values.	Essential to the framework is the maintenance of exhaustive documentation. This encompasses detailed records of the AI's architecture, the data used in training, and the processes underpinning its decision-making. The goal is to foster transparency and accountability throughout the entire lifecycle of the AI, bolstered by clearly defined responsibility protocols.	Central to the framework is the commitment to ensuring data rights and privacy. This involves the assurance of obtaining informed consent for the use of data, particularly when it pertains to personal or sensitive information. The framework mandates robust methodologies for data anonymisation to mitigate risks associated with the use of identifiable personal data.	The framework requires the implementation of robust verification systems to ascertain the authenticity of AI-generated content, especially in critical domains such as journalism and academic research. Techniques such as watermarking are employed to embed discernible signatures into AI outputs, thus denoting their artificial origin.	The framework mandates systematic assessments to identify potential vulnerabilities within generative AI systems. It emphasises the continuous refinement of AI models to safeguard against emerging threats, thereby ensuring their resilience and security	An integral component of the framework is the execution of fairness audits. These audits aim to uncover and rectify any biases present within the training data and the outputs of the AI models. Feedback mechanisms are instituted to refine models, addressing inadvertent biases or discriminatory outcomes.	The framework necessitates a thorough evaluation of the economic and social ramifications, with particular attention to potential job displacements. It advocates for the development of strategies to facilitate smooth transitions within the labour market, considering the ethical implications linked to various corporate structures.	The framework underscores the importance of public education regarding the diverse applications of generative AI and associated ethical concerns. It establishes mechanisms for garnering and integrating public feedback on the deployment of generative AI technologies.

© CohesionX (PTY) LTD

Cohesion X GenAi Standards	Articulation of Purpose and Contextual Precision	Transparency and Accountability Through Comprehensive Documentation	Data Rights and Privacy	Implementation of Authenticity Protocols	Ensuring Security and Robustness	Conducting Fairness Audits	Economic and Social Impact Assessment	Promotion of Public Engagement and Collaborative Efforts
LLM and GenAi Risk Factors	Misalignment with Intended Purpose: These AI systems may not always align their outputs with the intended purpose of the user, leading to results that are irrelevant, inappropriate, or even harmful, especially in sensitive contexts. Overgeneralization and Stereotyping: AI models, particularly those trained on large and diverse datasets, can sometimes resort to overgeneralisation, leading to stereotypical or biased outputs that do not suit the specific context or purpose of the task. Lack of Contextual Understanding: Gena-AI and LLMs can struggle to fully grasp the context of a request or conversation. This lack of deep contextual understanding can lead to responses	Difficulty in Understanding AI Decision-making: Without comprehensive documentation, users and stakeholders may find it challenging to understand how the AI makes decisions. This lack of understanding can lead to mistrust or misuse of the technology. Inability to Identify and Address Biases: Without detailed documentation, it becomes difficult to identify biases in AI models. This lack of visibility can allow biases to go unchecked, potentially leading to unfair or discriminatory outcomes. Challenges in Ensuring Compliance: In regulatory environments, comprehensive documentation is often required to demonstrate compliance with legal and ethical standards. A lack of such documentation	Unauthorised Data Access and Breaches: The risk of unauthorised access to personal or sensitive data is significant, especially if AI systems are not adequately secured. Data breaches can lead to privacy violations and the potential misuse of personal information. Inadequate Consent and Transparency: Often, users may not be fully aware or may not have explicitly consented to the extent of data collection and usage by AI systems. This lack of transparency and informed consent can lead to violations of data rights. Bias and Discrimination in Data: AI	Creation of Misleading or False Information: Gena-AI and LLMs can generate convincing but false or misleading content, contributing to misinformation and potentially causing confusion or harm. Difficulty in Verifying Source Authenticity: The ability of these AI systems to generate realistic content can make it challenging to distinguish between AI-generated and human-generated content, complicating efforts to verify the authenticity of information. Potential for Deepfakes and Impersonation: Gena-AI, in particular, can be used to create deepfakes or impersonate individuals in text, audio, or video formats, posing significant risks to personal privacy and public trust.	Vulnerability to Adversarial Attacks: These AI systems can be susceptible to adversarial attacks, where small, carefully crafted changes to input data can lead to incorrect outputs, compromising the integrity and reliability of the system. Data Privacy Breaches: Gena-AI and LLMs often require access to large datasets, which may include sensitive or personal information. There is a risk of data breaches, where this information could be exposed or misused. Model Theft and Replication: There is a risk of proprietary AI models being stolen or replicated. This not only	Bias in Training Data: If the data used to train these AI models is biased or unrepresentative, it can lead to biased outputs, perpetuating and amplifying existing societal or systemic biases. Inequitable Impact Across Different Groups: These technologies might not perform equally well for different demographic groups, potentially leading to inequitable outcomes, such as less accuracy or relevance for certain populations. Opaque Decision-making Processes: The "black box" nature of many AI models makes it difficult to understand how decisions are made, complicating efforts to identify and address fairness issues.	Job Displacement and Labor Market Disruption: These technologies can automate tasks previously done by humans, potentially leading to job displacement and disruption in various sectors, raising concerns about unemployment and economic inequality. Weakening of Intellectual Property Rights: Gena-AI and LLMs can produce content that may infringe on existing intellectual property rights, leading to challenges in determining authorship and ownership and potentially devaluing original human-created content. Inequality in Access and Benefits: The advantages of these technologies might be	Reduced Trust and Acceptance: A lack of collaboration and transparency can significantly erode public trust and acceptance of AI technologies as misunderstandings or fears about their operation and impact grow. Misinterpretation of AI Capabilities: Without proper public education, there is a high risk of end-users developing unrealistic expectations or misunderstandings about what these AI systems can and cannot do, potentially leading to misapplication or overreliance. Inadequate Consideration of Ethical and Societal Implications: The absence of interdisciplinary collaboration can result in AI systems that fail to consider or address vital ethical and societal implications,

© CohesionX (PTY) LTD

	or content generation that misses nuances or is contextually inappropriate.	can result in non-compliance risks.	systems can inherit biases present in their training data, leading to discriminatory outcomes. If personal data reflects historical or societal biases, AI technologies can perpetuate and amplify these.		presents a direct economic threat to the creators but also raises concerns about using these models unethically.		disproportionately available to those with more resources or technical expertise, exacerbating existing social and economic inequalities.	potentially causing harm or societal disruption.
Cohesion X GenAi Standards	**Articulation of Purpose and Contextual Precision**	**Transparency and Accountability Through Comprehensive Documentation**	**Data Rights and Privacy**	**Implementation of Authenticity Protocols**	**Ensuring Security and Robustness**	**Conducting Fairness Audits**	**Economic and Social Impact Assessment**	**Promotion of Public Engagement and Collaborative Efforts**

© CohesionX (PTY) LTD

LLM and GenAi Risk Factors	Difficulty in Handling Ambiguity: These AI systems can have difficulty dealing with ambiguous inputs, leading to outputs that may be misaligned with the user's intentions or expectations. Reliance on Training Data Quality: The precision and relevance of outputs from Gena-AI and LLMs are highly dependent on the quality and breadth of their training data. Inadequate or biased training data can lead to outputs that are not fit for the intended purpose or lack contextual accuracy.	Barriers to Effective Audit and Oversight: Insufficient documentation hinders effective auditing and oversight, as external reviewers or internal teams may struggle to thoroughly assess the AI system's performance, safety, and fairness. Impediments to Knowledge Transfer and Collaboration: When documentation is lacking, it becomes harder to share knowledge about the AI system's capabilities and limitations, impeding collaboration and hindering the transfer of knowledge both within and between organisations	Data Misuse and Exploitation: There is a risk that collected data, under the guise of improving services or AI performance, can be misused for purposes other than originally intended, such as targeted advertising, manipulation, or unauthorised sale of data. Compliance with Data Protection Regulations: Ensuring compliance with a growing and diverse set of data protection laws (like GDPR CCPA) is challenging. Non-compliance not only poses legal risks but also risks losing user trust and credibility.	Manipulation and Exploitation Risks: There's a risk that malicious actors could exploit these technologies to manipulate public opinion, forge fake documents, or create fraudulent content for scams. Intellectual Property and Authorship Issues: Determining authorship and intellectual property rights for AI-generated content can be complex, raising questions about originality and ownership, especially in cases where the AI's output is heavily based on pre-existing human-created content.	Manipulation of Output: There's a potential for these AI systems to be manipulated to generate false or harmful outputs, either through direct hacking or by feeding them misleading input data. Dependency and Single Point of Failure: Over-reliance on Gena-AI and LLMs for critical processes can create a single point of failure. A breach or failure in these AI systems can lead to significant disruptions and pose serious security risks.	Overgeneralization: Gena-AI and LLMs might generate outputs that overgeneralise based on limited or biased data, leading to stereotypical or overly simplified representations. Lack of Standards for Measuring Fairness: There is no universal standard or consensus on what constitutes fairness in AI, making it challenging to design, evaluate, and regulate these systems for fair outcomes.	Dependence and Reduced Human Skill Development: Over-reliance on AI for decision-making and creative processes could lead to a decline in certain human skills and expertise and increased dependence on automated systems. Ethical and Cultural Considerations: These technologies might inadvertently undermine cultural diversity and ethical norms by promoting homogenised, AI-generated content, potentially impacting cultural industries and ethical standards in society.	Barriers to Meaningful Feedback and Improvement: Lack of informed public engagement can create a significant gap in feedback necessary for AI improvement, limiting the development of AI tools that are truly user-centric and effective.

© CohesionX (PTY) LTD

Cohesion X GenAi Standards	Articulation of Purpose and Contextual Precision	Transparency and Accountability Through Comprehensive Documentation	Data Rights and Privacy	Implementation of Authenticity Protocols	Ensuring Security and Robustness	Conducting Fairness Audits	Economic and Social Impact Assessment	Promotion of Public Engagement and Collaborative Efforts
Guideline on Mitigation	Clear Articulation of Purpose: Ensuring that the purpose of the AI or LLM is clearly defined and communicated is crucial. This involves: Transparently stating the objectives and intended use of the AI system. Setting boundaries for its applications to prevent misuse or unintended consequences. Establishing guidelines for users to understand the scope and limitations of the technology. Contextual Precision in Responses and Actions: AI systems, especially LLMs, should be designed to understand and appropriately respond to context. This involves: Developing algorithms capable of discerning nuances in different contexts, languages, and cultures. Implementing	Ensuring Comprehensive and Accessible Documentation: - Develop detailed documentation for AI and LLM systems, explaining how they function, the nature of the data they use, and the decision-making processes involved. - Make this documentation accessible to relevant stakeholders, including users, developers, and regulators, to enhance understanding and accountability. Implementing Data Governance and Privacy Standards: - Adopt robust data governance frameworks ensuring ethical collection, use, and storage. - Ensure compliance with global data protection regulations like GDPR, HIPAA, or others pertinent to the operating region.	Adherence to Data Protection Regulations: Comply with international data protection laws such as the GDPR, CCPA, and others, ensuring all AI and LLM operations respect user privacy and data rights. Regularly update data handling practices to align with evolving legal standards and requirements. Data Minimization and Anonymization: Practice data minimisation by collecting only the strictly necessary data for the intended purpose. Implement data anonymisation techniques to remove or modify personal information, ensuring that individual users cannot be identified from the data used by AI and LLMs.	Digital Watermarking and Fingerprinting: Integrate digital watermarking techniques into AI outputs. This allows for tracing AI-generated content back to its source, helping to establish its authenticity. Use digital fingerprinting to uniquely identify and track AI-generated data, ensuring its origin and alterations can be audited. Blockchain for Traceability and Verification: Utilise blockchain technology to create a transparent and immutable record of AI transactions and interactions. This can be particularly effective in scenarios where the verification of data origin and process integrity is crucial.	Regular Security Audits and Vulnerability Assessments: Conduct thorough and frequent security audits to identify potential vulnerabilities in AI and LLM systems. Perform penetration testing and vulnerability assessments to proactively discover and address security weaknesses. Implementation of Advanced Encryption Techniques: Use state-of-the-art encryption methods to protect data at rest, in transit, and during processing by AI systems. Regularly update cryptographic protocols to stay ahead of evolving cyber threats.	Comprehensive Bias Detection and Analysis: Perform thorough analysis to detect biases in AI algorithms and the data sets on which they are trained. This involves identifying and measuring biases related to race, gender, ethnicity, or other relevant social factors. Use statistical methods and fairness metrics to assess and quantify the extent of biases present. Diverse and Inclusive Training Data: Ensure the training data for AI and LLMs is representative of diverse populations and scenarios. Diverse datasets help in reducing systemic biases and increase the fairness of outcomes. Regularly update and expand the training datasets to reflect real-world diversity and evolving social contexts.	Comprehensive Impact Studies and Research: Conduct in-depth studies to understand AI and LLM deployment's potential economic and social impacts. This includes analysing effects on employment, economic inequality, social interactions, and cultural dynamics. Utilise a multidisciplinary approach involving economists, sociologists, and other relevant experts. Stakeholder Engagement and Public Consultation: Engage with a wide range of stakeholders, including industry experts, policymakers, community leaders, and the general public, to gather diverse perspectives on the potential impacts of AI and LLMs. Organise public consultations and forums to discuss	Unbalanced Regulation and Policy Development: Without diverse public engagement and collaboration, the development of AI regulations and policies may not be comprehensive or balanced, potentially leading to ineffective or overly restrictive frameworks that don't adequately address the nuanced challenges of AI technologies.

© CohesionX (PTY) LTD

	mechanisms that allow the AI to seek clarification or abstain from responding when the context is unclear or outside its domain of reliability.	findings and gather feedback.

© CohesionX (PTY) LTD

Cohesion X GenAi Standards	Articulation of Purpose and Contextual Precision	Transparency and Accountability Through Comprehensive Documentation	Data Rights and Privacy	Implementation of Authenticity Protocols	Ensuring Security and Robustness	Conducting Fairness Audits	Economic and Social Impact Assessment	Promotion of Public Engagement and Collaborative Efforts
Guideline on Mitigation		Transparent Data Usage and Consent Mechanisms: - Implement clear consent mechanisms where users are informed about what data is collected, how it is used, and the purpose behind it. - Provide users with the option to opt in or out of data collection processes and make it easy to exercise these rights. Regular Auditing and Reporting for Transparency: - Conduct regular audits of AI and LLM systems to assess compliance with data privacy standards and documentation accuracy. - Publicly report audit findings to demonstrate accountability and transparency in operations. Developing and Integrating Explainable AI (XAI) Methods: - Invest in research and development of Explainable AI technologies that make AI decision-making processes more transparent and understandable. - Integrate XAI into AI and LLM systems to provide clear explanations of outputs, particularly in critical applications	Robust Consent Mechanisms: Establish clear and transparent consent mechanisms, allowing users to understand what data is collected and how it's used. Provide users with easy-to-use options to give, withdraw, or manage consent regarding their data. Regular Privacy Impact Assessments: Conduct regular privacy impact assessments to identify potential data handling and processing risks. Use the findings to refine privacy practices and mitigate identified risks proactively. Transparent Data Usage Policies: Develop and maintain clear, easily understandable data usage policies. Ensure these policies are accessible to users and stakeholders, detailing how data is collected, used, stored, and shared.	Implement smart contracts for automated compliance checks to ensure AI systems adhere to predefined authenticity protocols. AI Output Verification Systems: Develop and implement systems specifically designed to verify the authenticity of outputs produced by AI and LLMs. These systems can use a combination of signature matching, pattern recognition, and consistency checking. Regularly update verification systems to adapt to evolving AI capabilities and potential manipulations. Robust Metadata Management: Ensure that all AI-generated content and decisions are accompanied by comprehensive metadata that details the information's source, process, and nature. This aids in verifying authenticity and tracing content back to its AI origin.	Robust Access Control and Authentication Mechanisms: Implement strict access control policies and authentication procedures to ensure only authorised personnel can interact with AI systems. Utilise multi-factor authentication, role-based access control, and continuous monitoring of access patterns. AI-Specific Cybersecurity Training for Staff: Provide specialised cybersecurity training for staff working with AI and LLMs, focusing on these technologies' unique challenges and risks. Keep the team updated on the latest cyber threats targeting AI systems and best practices for defence. Secure Development Lifecycle (SDL) Integration: Incorporate security considerations at every stage of the AI development	Iterative Testing and Model Adjustments: Implement an iterative process where AI models are continuously tested and refined to address detected fairness issues. Adjust algorithms and retrain models with corrected or enhanced datasets to improve fairness. Stakeholder and Expert Involvement in Audits: Involve a broad range of stakeholders, including those from underrepresented groups, in the auditing process. This can provide diverse perspectives and insights into potential fairness issues. Collaborate with external experts in fields like ethics, sociology, and law to ensure comprehensive fairness evaluations. Transparent Reporting and Documentation: Maintain transparency by documenting and reporting the findings from fairness audits. This	Developing and Implementing Mitigation Strategies: Based on the impact assessments, develop strategies to mitigate negative impacts, such as job displacement or societal inequalities. This could involve re-skilling programs, educational initiatives, and policy recommendations. Implement these strategies in collaboration with relevant stakeholders, including government bodies, educational institutions, and NGOs. Regular Monitoring and Reporting: Establish a system for regular monitoring of the ongoing economic and social impacts of AI and LLMs. Report these findings transparently to stakeholders and the public, ensuring ongoing accountability. Inclusive Design and Development: Ensure the design and development	

© CohesionX (PTY) LTD

		where understanding AI decisions is vital.			lifecycle, from design to deployment and maintenance.	includes detailing identified biases, the methods used for detection, and steps taken to address them.	process of AI systems is inclusive, considering diverse population groups' needs and potential impacts.	
Cohesion X GenAi Standards	**Articulation of Purpose and Contextual Precision**	**Transparency and Accountability Through Comprehensive Documentation**	**Data Rights and Privacy**	**Implementation of Authenticity Protocols**	**Ensuring Security and Robustness**	**Conducting Fairness Audits**	**Economic and Social Impact Assessment**	**Promotion of Public Engagement and Collaborative Efforts**

© CohesionX (PTY) LTD

Guideline on Mitigation	Strengthening Accountability Mechanisms**: - Establish clear accountability frameworks that define who is responsible for the actions and decisions made by AI systems. - Include mechanisms for redressal in cases where AI systems violate data rights or privacy standards.	Implementing Data Encryption and Secure Storage: Use advanced encryption methods to protect data during transit and storage. Ensure secure storage solutions are employed to prevent unauthorised access to sensitive data. User Education and Transparency: Educate users about their data rights and the privacy practices of AI and LLM systems. Be transparent about the AI's decision-making processes, especially in cases where user data influences outcomes.	Implement standards for metadata to ensure consistency and reliability across different AI systems. Ethical and Legal Standards for AI-generated Content: Establish ethical guidelines and legal standards that mandate the disclosure of AI involvement in content creation, especially in sensitive fields like news, academic research, or legal documentation. Integrate these standards into the design and operation of AI systems to enforce authentic and transparent content generation.	Utilise secure coding practices and conduct security reviews and testing as integral parts of the development process. Incident Response Planning and Management: Develop and maintain a robust incident response plan specifically tailored to AI systems, ensuring quick and effective action in case of a cybersecurity breach. Regularly test and update the incident response plan to stay prepared for new types of cyberattacks. Continual Monitoring and Machine Learning for Threat Detection: AI systems are continuously monitored to detect unusual activities or potential security breaches. Implement machine learning algorithms to quickly identify and respond to novel threats and anomalies.	Publicly share these reports to maintain accountability and foster trust among users and stakeholders. Developing Fairness Guidelines and Standards: Establish clear guidelines and standards for fairness in AI and LLMs within the organisation. Ensure these standards are aligned with ethical principles and relevant legal frameworks concerning anti-discrimination and fairness. Ongoing Monitoring and Feedback Loops: Set up mechanisms for continuous monitoring of AI performance to quickly identify and address any emerging fairness issues. Implement feedback loops that allow users to report perceived biases or unfair outcomes, facilitating ongoing improvements.	Involve representatives from different socio-economic backgrounds in the design process to ensure their needs and perspectives are considered. Policy Advocacy and Collaboration with Regulators: Collaborate with policymakers to advocate for regulations and policies that address AI and LLMs' potential economic and social impacts. Work towards creating a regulatory environment that supports positive outcomes while mitigating risks. Creating Economic and Social Value: Focus on developing AI applications that create social and economic value, such as improving healthcare, education, and environmental sustainability. Invest in AI initiatives that aim to solve societal challenges and contribute to economic growth in a sustainable manner.

© CohesionX (PTY) LTD

Specific Risk Mitigation Pathways	**Clear Articulation of Purpose:** Transparently state the objectives and intended use of the AI system to avoid misalignments with its purpose. **Setting Boundaries for Applications:** Establish clear boundaries for AI applications to prevent misuse or unintended consequences. **Guidelines for User Understanding:** Provide users with clear guidelines to understand the scope and limitations of the AI technology they are interacting with. **Contextual Analysis Algorithms:** Develop algorithms capable of discerning nuances in different contexts, languages, and cultures to ensure contextual precision in responses. **Mechanisms for Context Clarification:** Implement mechanisms that allow the AI system to seek clarification or abstain from responding when the context is unclear. **Human Oversight for Validation:** Introduce human-in-the-loop validation mechanisms to assess the appropriateness and relevance of AI-generated outputs, especially in sensitive contexts. **Feedback Loops for Continuous Improvement:** Incorporate feedback loops that allow users to provide input on the relevance and appropriateness of AI-generated outputs. **Ethical Guidelines and Compliance:** Establish ethical guidelines and compliance measures to ensure responsible use of AI and adherence to ethical standards. **Data Governance Frameworks:** Adopt robust data governance frameworks to ensure ethical collection, use, and storage of data. **Comprehensive Documentation:** Develop detailed documentation for AI systems, explaining their function, data usage, and decision-making processes. **Regular Auditing for Transparency:** Conduct regular audits of AI systems to assess compliance with data privacy standards and documentation accuracy. **Explainable AI Integration:** Invest in the development of Explainable AI (XAI) technologies to make AI decision-making processes transparent and understandable. **Stakeholder Engagement:** Actively involve diverse stakeholders in decision-making processes related to AI development to promote inclusivity and transparency. **Public Awareness Campaigns:** Launch public awareness campaigns to educate individuals about AI technologies and their societal impacts. **Participatory Design Processes:** Employ participatory design processes involving end-users and affected communities in developing AI applications. **Multi-Stakeholder Partnerships:** Facilitate multi-stakeholder partnerships to address complex AI-related challenges through collaboration and coordinated action. **Adherence to Data Protection Regulations:** Comply with international data protection laws such as GDPR and CCPA, ensuring AI operations respect user privacy and data rights. **Data Minimization and Anonymization:** Practice data minimisation and implement data anonymisation techniques to protect user privacy. **Transparent Data Usage Policies:** Develop clear data usage policies that detail how data is collected, used, stored, and shared. **Regular Security Audits:** Conduct regular security audits to identify and address potential vulnerabilities in AI systems, ensuring robust cybersecurity measures.

© CohesionX (PTY) LTD

The proposed framework, rooted in AI ethics principles drawn from the European Union's guidance, applies these principles across diverse domains to address risks associated with Large Language Models (LLMs) in generative AI applications. Emphasising ethical concerns, the framework identifies specific risk factors and proposes strategies for their mitigation, aiming for responsible AI deployment.

Governance strategies proposed alongside the framework facilitate ethical decision-making and oversight throughout the AI lifecycle. While focusing on LLMs, the framework complements existing risk assessment guidelines without supplanting them.

It catalyses discussions and refinements toward broader adoption and integration into best practices, fostering collaboration to ensure ethically sound AI products. The framework aims to deliver value while minimising risks across various domains by encompassing LLMs within generative AI.

Adaptability and Continuous Evolution

In recognition of the dynamic nature of the field, the Cohesion X Framework is deliberately designed to be adaptable. It undergoes periodic reviews to incorporate emerging issues, challenges, and innovations, thus ensuring its continued relevance and effectiveness in guiding ethical AI development

Framework Principles

1. Human Agency and Oversight
2. Technical Robustness and safety
3. Privacy and Data Protection
4. Transparency
5. Promotion of Diversity, Non-Discrimination, and Fairness
6. Societal and Environmental Well-being
7. Accountability

© CohesionX (PTY) LTD

5.4. Framework Detail

I. Articulation of Purpose and Contextual Precision

Description:

Paramount to the framework is the establishment of a lucid and precise articulation of the intended purpose behind each generative AI system. This encompasses a deep understanding of such systems' varied contexts and applications, ensuring alignment with the established societal norms and values.

Risk Factors:

- Misalignment with Intended Purpose: These AI systems may not always align their outputs with the intended purpose of the user, leading to results that are irrelevant, inappropriate, or even harmful, especially in sensitive contexts.

- Lack of Contextual Understanding: Gena-AI and LLMs can struggle to grasp the context of a request or conversation fully. This lack of deep contextual understanding can lead to responses or content generation that misses nuances or is contextually inappropriate.

Mitigation Pathways

- Clear Articulation of Purpose: Ensuring that the purpose of the AI or LLM is clearly defined and communicated is crucial. This involves:
    - Transparently stating the objectives and intended use of the AI system.
    - Setting boundaries for its applications to prevent misuse or unintended consequences.
    - Establishing guidelines for users to understand the scope and limitations of the technology.

- Contextual Precision in Responses and Actions: AI systems, especially LLMs, should be designed to understand and appropriately respond to context. This involves:

- Developing algorithms capable of discerning nuances in different contexts, languages, and cultures.

- Implementing mechanisms that allow the AI to seek clarification or abstain from responding when the context is unclear or outside its domain of reliability.

II. Transparency and Accountability Through Comprehensive Documentation

Description

Essential to the framework is the maintenance of exhaustive documentation. This encompasses detailed records of the AI's architecture, the data used in training, and the processes underpinning its decision-making. The goal is to foster transparency and accountability throughout the entire lifecycle of the AI, bolstered by clearly defined responsibility protocols.

Risk Factors:

© CohesionX (PTY) LTD

- Difficulty in Understanding AI Decision-making: Without comprehensive documentation, users and stakeholders may find it challenging to understand how the AI makes decisions. This lack of understanding can lead to mistrust or misuse of the technology.

- Inability to Identify and Address Biases: Without detailed documentation, it becomes challenging to identify biases in AI models. This lack of visibility can allow biases to go unchecked, potentially leading to unfair or discriminatory outcomes.

- Challenges in Ensuring Compliance: In regulatory environments, comprehensive documentation is often required to demonstrate compliance with legal and ethical standards. A lack of such documentation can result in non-compliance risks.

- Barriers to Effective Audit and Oversight: Insufficient documentation hinders effective auditing and oversight, as external reviewers or internal teams may struggle to thoroughly assess the AI system's performance, safety, and fairness.

- Impediments to Knowledge Transfer and Collaboration: When documentation is lacking, it becomes harder to share knowledge about the AI system's capabilities and limitations, impeding collaboration and hindering the transfer of knowledge both within and between organisations

Mitigation Pathways

- Ensuring Comprehensive and Accessible Documentation:
  Develop detailed documentation for AI and LLM systems, explaining how they function, the nature of the data they use, and the decision-making processes involved.
  Make this documentation accessible to relevant stakeholders, including users, developers, and regulators, to enhance understanding and accountability.

- Implementing Data Governance and Privacy Standards:
  Adopt robust data governance frameworks that ensure ethical data collection, use, and storage.
  Ensure compliance with global data protection regulations like GDPR, HIPAA, or others pertinent to the operating region.

- Transparent Data Usage and Consent Mechanisms:
  Implement clear consent mechanisms where users are informed about what data is collected, how it is used, and the purpose behind it.
  Provide users with the option to opt in or out of data collection processes and make it easy to exercise these rights.

- Regular Auditing and Reporting for Transparency:
  Conduct regular audits of AI and LLM systems to assess compliance with data privacy standards and documentation accuracy.
  Publicly report audit findings to demonstrate accountability and transparency in operations.

- Developing and Integrating Explainable AI (XAI) Methods:
  Invest in research and development of Explainable AI technologies that make AI decision-making processes more transparent and understandable.
  Integrate XAI into AI and LLM systems to explain outputs clearly, particularly in critical applications where understanding AI decisions is vital.

- Strengthening Accountability Mechanisms:
  Establish clear accountability frameworks that define who is responsible for the actions and decisions made by AI systems.

© CohesionX (PTY) LTD

Include mechanisms for redressal in cases where AI systems violate data rights or privacy standards.

III. Data Rights and Privacy

Description:

Central to the framework is the commitment to ensuring data rights and privacy. This involves the assurance of obtaining informed consent for the use of data, particularly when it pertains to personal or sensitive information. The framework mandates robust methodologies for data anonymisation to mitigate risks associated with the use of identifiable personal data.

Risk Factors:

- Unauthorised Data Access and Breaches: The risk of unauthorised access to personal or sensitive data is significant, especially if AI systems are not adequately secured. Data breaches can lead to privacy violations and the potential misuse of personal information.
- Inadequate Consent and Transparency: Often, users may not be fully aware or may not have explicitly consented to the extent of data collection and usage by AI systems. This lack of transparency and informed consent can lead to violations of data rights.
- Bias and Discrimination in Data: AI systems can inherit biases present in their training data, leading to discriminatory outcomes. If personal data reflects historical or societal biases, AI technologies can perpetuate and amplify these.
- Data Misuse and Exploitation: There is a risk that collected data, under the guise of improving services or AI performance, can be misused for purposes other than initially intended, such as targeted advertising, manipulation, or unauthorised sale of data.
- Compliance with Data Protection Regulations: Ensuring compliance with a growing and diverse set of data protection laws (like GDPR CCPA) is challenging. Non-compliance not only poses legal risks but also risks losing user trust and credibility.

Mitigation Pathways

- Adherence to Data Protection Regulations:
  Comply with international data protection laws such as the GDPR, CCPA, and others, ensuring all AI and LLM operations respect user privacy and data rights.
  Regularly update data handling practices to align with evolving legal standards and requirements.
- Data Minimization and Anonymization:
  Practice data minimisation by collecting only the strictly necessary data for the intended purpose.
  Implement data anonymisation techniques to remove or modify personal information, ensuring that individual users cannot be identified from the data used by AI and LLMs.
- Robust Consent Mechanisms:
  Establish clear and transparent consent mechanisms, allowing users to understand what data is collected and how it's used.
  Provide users with easy-to-use options to give, withdraw, or manage consent regarding their data.
  Regular Privacy Impact Assessments:

© CohesionX (PTY) LTD

- Conduct regular privacy impact assessments to identify potential data handling and processing risks.
  Use the findings to refine privacy practices and mitigate identified risks proactively.

- Transparent Data Usage Policies:
  Develop and maintain clear, easily understandable data usage policies.
  Ensure these policies are accessible to users and stakeholders, detailing how data is collected, used, stored, and shared.

- Implementing Data Encryption and Secure Storage:
  Use advanced encryption methods to protect data during transit and storage.
  Ensure secure storage solutions are employed to prevent unauthorised access to sensitive data.

- User Education and Transparency:
  Educate users about their data rights and the privacy practices of AI and LLM systems.
  Be transparent about the AI's decision-making processes, especially in cases where user data influences outcomes.

IV. Implementation of Authenticity Protocols:

Description

The framework requires the implementation of robust verification systems to ascertain the authenticity of AI-generated content, especially in critical domains such as journalism and academic research. Techniques such as watermarking are employed to embed discernible signatures into AI outputs, thus denoting their artificial origin.

Risk Factors:

- Creation of Misleading or False Information: Gena-AI and LLMs can generate convincing but false or misleading content, contributing to misinformation and potentially causing confusion or harm.

- Difficulty in Verifying Source Authenticity: The ability of these AI systems to generate realistic content can make it challenging to distinguish between AI-generated and human-generated content, complicating efforts to verify the authenticity of information.

- Potential for Deepfakes and Impersonation: Gena-AI, in particular, can be used to create deepfakes or impersonate individuals in text, audio, or video formats, posing significant risks to personal privacy and public trust.

- Manipulation and Exploitation Risks: There's a risk that malicious actors could exploit these technologies to manipulate public opinion, forge fake documents, or create fraudulent content for scams.

- Intellectual Property and Authorship Issues: Determining authorship and intellectual property rights for AI-generated content can be complex, raising questions about originality and ownership, especially in cases where the AI's output is heavily based on pre-existing human-created content.

Mitigation Pathways

- Digital Watermarking and Fingerprinting:

© CohesionX (PTY) LTD

Integrate digital watermarking techniques into AI outputs. This allows for tracing AI-generated content back to its source, helping to establish its authenticity.
Use digital fingerprinting to uniquely identify and track AI-generated data, ensuring its origin and alterations can be audited.

- Blockchain for Traceability and Verification:
  Utilise blockchain technology to create a transparent and immutable record of AI transactions and interactions. This can be particularly effective in scenarios where the verification of data origin and process integrity is crucial.
  Implement smart contracts for automated compliance checks to ensure AI systems adhere to predefined authenticity protocols.

- AI Output Verification Systems:
  Develop and implement systems specifically designed to verify the authenticity of outputs produced by AI and LLMs. These systems can use a combination of signature matching, pattern recognition, and consistency checking.
  Regularly update verification systems to adapt to evolving AI capabilities and potential manipulations.

- Robust Metadata Management:
  Ensure that all AI-generated content and decisions are accompanied by comprehensive metadata that details the information's source, process, and nature. This aids in verifying authenticity and tracing content back to its AI origin.
  Implement standards for metadata to ensure consistency and reliability across different AI systems.

- Ethical and Legal Standards for AI-generated Content:
  Establish ethical guidelines and legal standards that mandate the disclosure of AI involvement in content creation, especially in sensitive fields like news, academic research, or legal documentation.
  Integrate these standards into the design and operation of AI systems to enforce authentic and transparent content generation.
  AI and LLM developers can significantly bolster the authenticity of AI-generated content and decisions by employing these pathways. This not only mitigates risks related to authenticity but also enhances trust and reliability in AI technologies.

V. Ensuring Security and Robustness:

Description

The framework mandates systematic assessments to identify potential vulnerabilities within generative AI systems. It emphasises the continuous refinement of AI models to safeguard against emerging threats, thereby ensuring their resilience and security.

Risk Factors

- Vulnerability to Adversarial Attacks: These AI systems can be susceptible to adversarial attacks, where small, carefully crafted changes to input data can lead to incorrect outputs, compromising the integrity and reliability of the system.

- Data Privacy Breaches: Gena-AI and LLMs often require access to large datasets, which may include sensitive or personal information. There is a risk of data breaches, where this information could be exposed or misused.

© CohesionX (PTY) LTD

- Model Theft and Replication: There is a risk of proprietary AI models being stolen or replicated. This not only presents a direct economic threat to the creators but also raises concerns about using these models unethically.

- Manipulation of Output: There's a potential for these AI systems to be manipulated to generate false or harmful outputs, either through direct hacking or by feeding them misleading input data.

- Dependency and Single Point of Failure: Over-reliance on Gena-AI and LLMs for critical processes can create a single point of failure. A breach or failure in these AI systems can lead to significant disruptions and pose serious security risks.

Mitigation Pathways

- Regular Security Audits and Vulnerability Assessments:
  Conduct thorough and frequent security audits to identify potential vulnerabilities in AI and LLM systems.
  Perform penetration testing and vulnerability assessments to discover and address security weaknesses proactively.

- Implementation of Advanced Encryption Techniques:
  Use state-of-the-art encryption methods to protect data at rest, in transit, and during processing by AI systems.
  Regularly update cryptographic protocols to stay ahead of evolving cyber threats.

- Robust Access Control and Authentication Mechanisms:
  Implement strict access control policies and authentication procedures to ensure only authorised personnel can interact with AI systems.
  Utilise multi-factor authentication, role-based access control, and continuous monitoring of access patterns.

- AI-Specific Cybersecurity Training for Staff:
  Provide specialised cybersecurity training for staff working with AI and LLMs, focusing on these technologies' unique challenges and risks.
  Keep the team updated on the latest cyber threats targeting AI systems and best practices for defence.

- Secure Development Lifecycle (SDL) Integration:
  Incorporate security considerations at every stage of the AI development lifecycle, from design to deployment and maintenance.
  Utilise secure coding practices and conduct security reviews and testing as integral parts of the development process.

- Incident Response Planning and Management:
  Develop and maintain a robust incident response plan tailored explicitly to AI systems, ensuring quick and effective action in case of a cybersecurity breach.
  Regularly test and update the incident response plan to stay prepared for new types of cyberattacks.

- Continual Monitoring and Machine Learning for Threat Detection:
  AI systems are continuously monitored to detect unusual activities or potential security breaches.
  Implement machine learning algorithms to identify and respond to novel threats and anomalies quickly.

© CohesionX (PTY) LTD

By focusing on these pathways, organisations can greatly enhance the cybersecurity and robustness of their AI and LLM systems, reducing the risks associated with digital threats and vulnerabilities.

VI. Conducting Fairness Audits:

Description

An integral component of the framework is the execution of fairness audits. These audits aim to uncover and rectify any biases present within the training data and the outputs of the AI models. Feedback mechanisms are instituted to refine models, addressing inadvertent biases or discriminatory outcomes.

Risk Factors:

- Bias in Training Data: If the data used to train these AI models is biased or unrepresentative, it can lead to biased outputs, perpetuating and amplifying existing societal or systemic biases.

- Inequitable Impact Across Different Groups: These technologies might not perform equally well for different demographic groups, potentially leading to inequitable outcomes, such as less accuracy or relevance for specific populations.

- Opaque Decision-making Processes: The "black box" nature of many AI models makes it difficult to understand how decisions are made, complicating efforts to identify and address fairness issues.

- Overgeneralization: Gena-AI and LLMs might generate outputs that overgeneralise based on limited or biased data, leading to stereotypical or overly simplified representations.

- Lack of Standards for Measuring Fairness: There is no universal standard or consensus on what constitutes fairness in AI, making it challenging to design, evaluate, and regulate these systems for fair outcomes.

Mitigation Pathways:

- Comprehensive Bias Detection and Analysis:
  Perform thorough analysis to detect biases in AI algorithms and the data sets on which they are trained. This involves identifying and measuring biases related to race, gender, ethnicity, or other relevant social factors.
  Use statistical methods and fairness metrics to assess and quantify the extent of biases present.

- Diverse and Inclusive Training Data:
  Ensure the training data for AI and LLMs is representative of diverse populations and scenarios. Diverse datasets help in reducing systemic biases and increase the fairness of outcomes.
  Regularly update and expand the training datasets to reflect real-world diversity and evolving social contexts.

- Iterative Testing and Model Adjustments:
  Implement an iterative process where AI models are continuously tested and refined to address detected fairness issues.

© CohesionX (PTY) LTD

Adjust algorithms and retrain models with corrected or enhanced datasets to improve fairness.

- Stakeholder and Expert Involvement in Audits:
  Involve a broad range of stakeholders, including those from underrepresented groups, in the auditing process. This can provide diverse perspectives and insights into potential fairness issues.
  Collaborate with external experts in fields like ethics, sociology, and law to ensure comprehensive fairness evaluations.
- Transparent Reporting and Documentation:
  Maintain transparency by documenting and reporting the findings from fairness audits. This includes detailing identified biases, the methods used for detection, and steps taken to address them.
  Publicly share these reports to maintain accountability and foster trust among users and stakeholders.

- Developing Fairness Guidelines and Standards:
  Establish clear guidelines and standards for fairness in AI and LLMs within the organisation. Ensure these standards are aligned with ethical principles and relevant legal frameworks concerning anti-discrimination and fairness.

- Ongoing Monitoring and Feedback Loops:
  Set up mechanisms for continuous monitoring of AI performance to quickly identify and address any emerging fairness issues.
  Implement feedback loops that allow users to report perceived biases or unfair outcomes, facilitating ongoing improvements.

VII. Economic and Social Impact Assessment:

Description

The framework necessitates a thorough evaluation of the economic and social ramifications, with particular attention to potential job displacements. It advocates for the development of strategies to facilitate smooth transitions within the labour market, considering the ethical implications linked to various corporate structures.

Risk Factors

- Job Displacement and Labour Market Disruption: These technologies can automate tasks previously done by humans, potentially leading to job displacement and disruption in various sectors, raising concerns about unemployment and economic inequality.

- Weakening of Intellectual Property Rights: Gena-AI and LLMs can produce content that may infringe on existing intellectual property rights, leading to challenges in determining authorship and ownership and potentially devaluing original human-created content.

- Inequality in Access and Benefits: The advantages of these technologies might be disproportionately available to those with more resources or technical expertise, exacerbating existing social and economic inequalities.

- Dependence and Reduced Human Skill Development: Over-reliance on AI for decision-making and creative processes could lead to a decline in specific human skills and expertise and increased dependence on automated systems.

© CohesionX (PTY) LTD

- Ethical and Cultural Considerations: These technologies might inadvertently undermine cultural diversity and ethical norms by promoting homogenised, AI-generated content, potentially impacting cultural industries and ethical standards in society.

Mitigation Pathways

- Comprehensive Impact Studies and Research:
  Conduct in-depth studies to understand AI and LLM deployment's potential economic and social impacts. This includes analysing effects on employment, economic inequality, social interactions, and cultural dynamics.
  Utilise a multidisciplinary approach involving economists, sociologists, and other relevant experts.

- Stakeholder Engagement and Public Consultation:
  Engage with a wide range of stakeholders, including industry experts, policymakers, community leaders, and the general public, to gather diverse perspectives on the potential impacts of AI and LLMs.
  Organise public consultations and forums to discuss findings and gather feedback.

- Developing and Implementing Mitigation Strategies:
  Based on the impact assessments, develop strategies to mitigate negative impacts, such as job displacement or societal inequalities. This could involve re-skilling programs, educational initiatives, and policy recommendations.
  Implement these strategies in collaboration with relevant stakeholders, including government bodies, educational institutions, and NGOs.

- Regular Monitoring and Reporting:
  Establish a system for regular monitoring of the ongoing economic and social impacts of AI and LLMs.
  Report these findings transparently to stakeholders and the public, ensuring ongoing accountability.

- Inclusive Design and Development:
  Ensure the design and development process of AI systems is inclusive, considering diverse population groups' needs and potential impacts.
  Involve representatives from different socio-economic backgrounds in the design process to ensure their needs and perspectives are considered.

- Policy Advocacy and Collaboration with Regulators:
  Collaborate with policymakers to advocate for regulations and policies that address AI and LLMs' potential economic and social impacts.
  Work towards creating a regulatory environment that supports positive outcomes while mitigating risks.

- Creating Economic and Social Value:
  Focus on developing AI applications that create social and economic value, such as improving healthcare, education, and environmental sustainability.
  Invest in AI initiatives that aim to solve societal challenges and contribute to economic growth in a sustainable manner.

© CohesionX (PTY) LTD

VIII. Promotion of Public Engagement and Collaborative Efforts:

Description

The framework underscores the importance of public education regarding the diverse applications of generative AI and associated ethical concerns. It establishes mechanisms for garnering and integrating public feedback on the deployment of generative AI technologies.

Risk Factors

- Reduced Trust and Acceptance: A lack of collaboration and transparency can significantly erode public trust and acceptance of AI technologies as misunderstandings or fears about their operation and impact grow.

- Misinterpretation of AI Capabilities: Without proper public education, there is a high risk of end-users developing unrealistic expectations or misunderstandings about what these AI systems can and cannot do, potentially leading to misapplication or overreliance.

- Inadequate Consideration of Ethical and Societal Implications: The absence of interdisciplinary collaboration can result in AI systems that fail to consider or address vital ethical and societal implications, potentially causing harm or societal disruption.

- Barriers to Meaningful Feedback and Improvement: Lack of informed public engagement can create a significant gap in feedback necessary for AI improvement, limiting the development of AI tools that are truly user-centric and effective.

- Unbalanced Regulation and Policy Development: Without diverse public engagement and collaboration, the development of AI regulations and policies may not be comprehensive or balanced, potentially leading to ineffective or overly restrictive frameworks that don't adequately address the nuanced challenges of AI technologies.

Mitigation Pathways

- Open Access Platforms and Forums: Establishing open-access platforms and forums dedicated to discussing AI development, its implications, and potential risks can encourage public participation. These platforms can serve as spaces for sharing information, gathering feedback, and fostering collaboration between AI developers, researchers, policymakers, and the general public.

- Public Awareness Campaigns: Launching public awareness campaigns to educate individuals about AI technologies, their capabilities, limitations, and potential societal impacts is crucial. These campaigns can involve workshops, seminars, online tutorials, and educational materials designed to empower people to engage meaningfully in discussions surrounding AI development.

- Stakeholder Consultations and Inclusivity: Actively involving diverse stakeholders, including community representatives, advocacy groups, ethicists, policymakers, and industry experts, in decision-making processes related to AI development can promote inclusivity and ensure that a wide range of perspectives are considered. This approach fosters collaborative efforts and helps address concerns raised by different stakeholders.

- Ethical Guidelines and Frameworks: Developing and adhering to robust ethical guidelines and frameworks for AI research and development can provide a roadmap for responsible innovation. These guidelines should prioritise transparency, fairness, accountability, and the

© CohesionX (PTY) LTD

protection of human rights. Engaging with the public and stakeholders during the formulation of such guidelines can enhance their legitimacy and effectiveness.

- Citizen Science Initiatives: Encouraging citizen science initiatives that involve the public in AI research projects can promote collaboration and knowledge sharing. Platforms that enable individuals to contribute data, insights, or feedback to AI projects under appropriate privacy and ethical considerations can foster a sense of ownership and responsibility among participants.

- Participatory Design Processes: Employing participatory design processes involving end-users and affected communities in developing AI applications can lead to more inclusive and user-centred solutions. Developers can better anticipate and address potential societal impacts and concerns by soliciting input from diverse stakeholders at various stages of the design and implementation process.

- Multi-Stakeholder Partnerships: Facilitating multi-stakeholder partnerships between governments, academia, civil society organisations, industry players, and international bodies can promote collaborative efforts to address complex AI-related challenges. These partnerships can facilitate information sharing, resource pooling, and coordinated action towards common goals, such as enhancing AI safety and governance.

Framework-Specific Mitigation Pathways

- Clear Articulation of Purpose:
  Transparently state the objectives and intended use of the AI system to avoid misalignments with its purpose.

- Setting Boundaries for Applications:
  Establish clear boundaries for the applications of AI to prevent misuse or unintended consequences.
- Guidelines for User Understanding:
  Provide users with clear guidelines to understand the scope and limitations of the AI technology they are interacting with.

- Contextual Analysis Algorithms:
  Develop algorithms capable of discerning nuances in different contexts, languages, and cultures to ensure contextual precision in responses.

- Mechanisms for Context Clarification:
  Implement mechanisms that allow the AI system to seek clarification or abstain from responding when the context is unclear.

- Human Oversight for Validation:
  Introduce human-in-the-loop validation mechanisms to assess the appropriateness and relevance of AI-generated outputs, especially in sensitive contexts.

- Feedback Loops for Continuous Improvement:
  Incorporate feedback loops that allow users to provide input on the relevance and appropriateness of AI-generated outputs.

- Ethical Guidelines and Compliance:

© CohesionX (PTY) LTD

Establish ethical guidelines and compliance measures to ensure responsible use of AI and adherence to ethical standards.

- Data Governance Frameworks:
  Adopt robust data governance frameworks to ensure ethical data collection, use, and storage.

- Comprehensive Documentation:
  Develop detailed documentation for AI systems, explaining their function, data usage, and decision-making processes.

- Regular Auditing for Transparency:
  Conduct regular audits of AI systems to assess compliance with data privacy standards and documentation accuracy.

- Explainable AI Integration:
  Invest in the development of Explainable AI (XAI) technologies to make AI decision-making processes transparent and understandable.

- Stakeholder Engagement:
  Actively involve diverse stakeholders in decision-making processes related to AI development to promote inclusivity and transparency.

- Public Awareness Campaigns:
  Launch public awareness campaigns to educate individuals about AI technologies and their societal impacts.

- Participatory Design Processes:
  Employ participatory design processes involving end-users and affected communities in developing AI applications.

- Multi-Stakeholder Partnerships:
  Facilitate multi-stakeholder partnerships to address complex AI-related challenges through collaboration and coordinated action.

- Adherence to Data Protection Regulations:
  Comply with international data protection laws such as GDPR CCPA, ensuring AI operations respect user privacy and data rights.

- Data Minimization and Anonymization:
  Practice data minimisation and implement data anonymisation techniques to protect user privacy.

- Transparent Data Usage Policies:
  Develop transparent data usage policies detailing data collection, storage, and sharing.

- Regular Security Audits:
  Conduct regular security audits to identify and address potential vulnerabilities in AI systems, ensuring robust cybersecurity measures.

Adaptability and Continuous Evolution

In recognition of the dynamic nature of the field, the Cohesion X Framework is deliberately designed to be adaptable. It undergoes periodic reviews to incorporate emerging issues,

© CohesionX (PTY) LTD

challenges, and innovations, thus ensuring its continued relevance and effectiveness in guiding ethical AI development.

Implementation of Framework

Our framework implementation, depicted in Figure 1 through a detailed block diagram, is an intricate system designed to address the ethical intricacies of generative artificial intelligence holistically. This framework is comprehensive in its approach and visually structured to facilitate ease of understanding and implementation.

1. Awareness and Ethical Issue Recognition:
   - This crucial first step involves the keen identification of ethical dilemmas specific to the realm of Generative AI.
   - Comprehensive Evaluation of Generative AI's Scope and Limitations. Focusing on a nuanced assessment of AI's capabilities and constraints ensures a grounded understanding of its practical implications and sphere of influence.

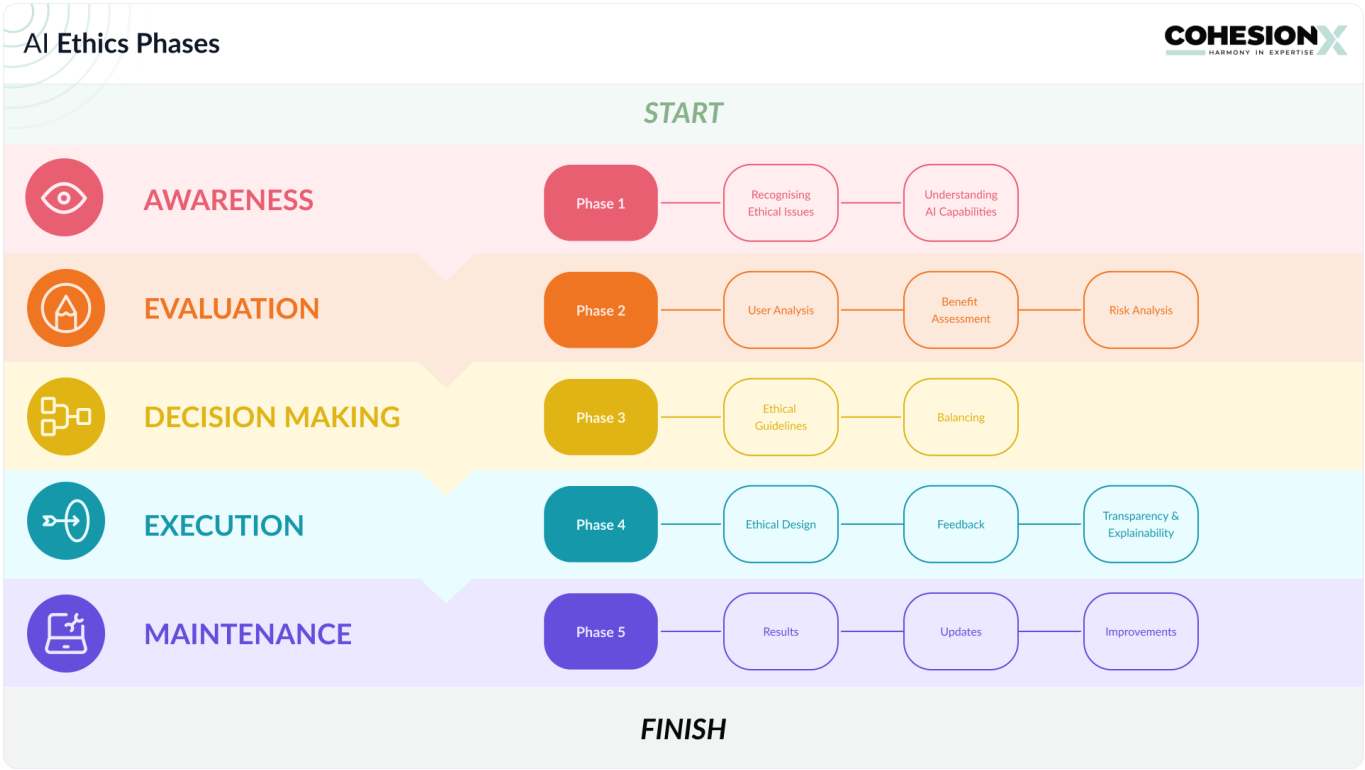
2. Evaluation and Stakeholder Analysis:
   - User and Impact Assessment: This involves a thorough examination of those affected by generative AI outputs, assessing the breadth and depth of its impact across diverse stakeholder groups.
   - Risk Analysis: This entails a critical evaluation of potential hazards, such as misinterpretation or misuse of AI-generated content, while gauging the probability and severity of these risks.
   - Benefit Assessment: A deliberate process to gauge the positive outcomes of generative AI, including its role in spurring creativity, enhancing decision-making, and contributing to societal well-being.

3. Decision-Making and Ethical Regulation:
   - Formulation of Ethical Guidelines: Crafting clear, actionable regulations grounded in universally recognised ethical principles. These guidelines aim to align the use of generative AI with societal norms of fairness, transparency, and confidentiality.
   - Equitable Balancing: A methodical process of weighing benefits against risks, considering various interests and viewpoints, to arrive at equitable and just decisions.

4. Implementation and Ethical Integration:
   - Ethics-Infused Design: Integrating ethical considerations into the very fabric of generative AI systems, promoting ethical usage from inception.
   - Transparency and Explainability in AI: Ensuring the AI's operations and decision-making processes are accessible and comprehensible to all stakeholders.
   - Stakeholder Feedback Mechanisms: Facilitating channels for stakeholders to voice their opinions on AI outputs and related ethical concerns, promoting a dynamic and responsive system.

5. Review and Iterative Refinement:
   - Ongoing Outcome Evaluation: Continuous monitoring of AI outputs to identify unforeseen ethical challenges or consequences that may arise over time.
   - Iterative Improvements: Adapting the AI system in response to feedback, evolving societal norms, and observed outcomes to maintain ethical and social relevance.
   - Regular Framework Updates: Keeping the framework abreast with the latest research and advancements in the field, ensuring its relevance and efficacy in addressing ethical complexities in generative AI.

© CohesionX (PTY) LTD



© CohesionX (PTY) LTD

Document References

This document references the following related documents stipulated below.

Ref No	Description
1	A Conceptual Framework for Solving Ethical Issues in Generative Artificial Intelligence Authors: Plamena Ventseslavova Zlateva, University of National and World Economy Liudmila Steshina, Igor Petukhov, Dimiter Velev, University of National and World Economy
2.	Attention is not all you need: the complicated case of ethically using large language models Author: Stefan Harrer, Digital Health Cooperative Research Centre, Melbourne, Australia
3.	EU guidelines on ethics in artificial intelligence: Context and implementation Author: Tambiama Madiega, EPRS | European Parliamentary Research Service

© CohesionX (PTY) LTD