ALIGNER D3.2

Risk Assessment of AI Technologies for EU LEAs

Deliverable No.	D3.2
Work Package	WP3
Dissemination Level	PU
Author(s)	Mathilde Jarlsbo (FOI), Norea Normelli (FOI), Peter Svenmarck (FOI) & Tomas Piatrik (CBRNE)
Co-Author(s)	-
Contributor(s)	Donatella Casaburo (KUL)
Due date	2023-12-31
Actual submission date	2024-05-24
Status	Final
Revision	1.1
Reviewed by (if applicable)	Monika Weinbuch (MPD), Daniel Lückerath (Fraunhofer)

This document has been prepared in the framework of the European project ALIGNER – Artificial Intelligence Roadmap for Policing and Law Enforcement. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101020574.

The sole responsibility for the content of this publication lies with the authors. It does not necessarily represent the opinion of the European Union. Neither the REA nor the European Commission are responsible for any use that may be made of the information contained therein.

Contact:

info@aligner-h2020.eu
www.aligner-h2020.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no. 101020574.

Executive Summary

One of the objectives of the European Commission-funded Coordination and Support Action ALIGNER (Artificial Intelligence Roadmap for Policing and Law Enforcement) is to identify promising Artificial Intelligence (AI) technologies and propose a roadmap for future research investments in AI for Law Enforcement Agencies (LEAs). The stakeholders in ALIGNER are European actors concerned with artificial intelligence (AI), law enforcement, and policing that collectively identify and discuss promising AI technologies for LEAs. Although AI technologies provide many benefits for LEAs and the society in general, they also contribute potential risks.

This report describes the ALIGNER risk assessment of AI technologies for LEAs to identify and mitigate potential risks. The ALIGNER Risk Assessment Instrument (RAI) is performed as a part of the ALIGNER AI Technology Watch method framework for impact assessment of AI technologies for LEAs (ALIGNER D3.1, Westman et al., 2022). The risk assessment instrument complements the AI technology impact assessment (ALIGNER D3.1, Westman et al., 2022) and the fundamental rights impact assessment (ALIGNER D4.2, Casaburo & Marsh, 2023).

After an introduction, the second section in this report includes a description of four already existing instruments for AI Technology Risk Assessment: Assessment List for Trustworthy Artificial Intelligence (ALTAI), Shaping the ethical dimensions of smart information systems – A European perspective (SHERPA), Algorithmic Impact Assessment (AIA) and Artificial Intelligence Toolkit (AIT). The third section complements the second one by introducing mitigation measures that may reduce the likelihood for or severity of risk realisation. The risks and related mitigation measures are categorised in five groups: Lawfulness, fairness and transparency of processing; Data and storage minimisation; Data accuracy and security; Data subject rights and access control; and Automated decision-making.

Next, the report describes the ALIGNER Risk Assessment Instrument (RAI) and its methodology in Section 4. The instrument aims to help LEAs identify risks related to AI technologies, assess the impact of those risks, and implement relevant mitigation measures to reduce the likelihood for or the severity of risk realisation. The instrument consist of seven templates to help LEAs consider a variety of relevant issues when determining the potential risks posed by use of AI, as well as helping LEAs to plan ways of responding to these risks. Finally, the report describes examples of how the ALIGNER RAI can be used together with the ALIGNER Scenario Cards drafted in the context of task 3.1. It is recommended that the instrument should be conducted periodically where interdisciplinary competence is very important. LEAs are responsible for conducting the ALIGNER Risk Assessment. It is recommended that it should be conducted by LEA personnel and supervisors who understand the technical, ethical, and legal issues. The ALIGNER RAI is not meant to replace the implementation of other risk assessment.

The ALIGNER RAI received strong support by LEA stakeholders at ALIGNER workshop no. 6 in November 2023. Participants only requested additional clarifications and details, which have been incorporated into this report.

Table of Contents

List of Abbreviations

Abbreviation	Meaning
ABAC	Attribute-Based Access Control
AES	Advanced Encryption Standard
AI	Artificial intelligence
AIA	Algorithmic impact assessment
AI HLEG	High-Level Expert Group on Artificial Intelligence
AIRA	Artificial Intelligence Risk Assessment
AIT	Artificial Intelligence Toolkit
ALIGNER	Artificial Intelligence Roadmap for Policing and Law Enforcement
ALTAI	Assessment List for Trustworthy Artificial Intelligence
DPO	Data Protection Officer
DL	Deep learning
DPIA	Data Protection Impact Assessment
EU	European Union
GDPR	General Data Protection Regulation
LEA	Law Enforcement Agency
LED	Law Enforcement Directive
MFA	Multi-Factor Authentication
RAI	Risk assessment instrument
RBAC	Role based access control
RSA	Rivest–Shamir–Adleman
SHERPA	Shaping the ethical dimensions of smart information systems – A European perspective
TLS	Transport Layer Security
UNICRI	United Nations Interregional Crime and Justice Research Institute

1. Introduction (FOI)

Scientific innovation that increases the capability to collect, store and process information from various sources provides many benefits for law enforcement agencies (LEAs). Much of this success is due to Artificial Intelligence (AI) technologies, such as deep-learning (DL) that trains models on a large volume of datasets to extract hidden patterns. Since there is a continuous development of AI technologies that potentially are useful for LEAs, it is important to identify the most promising technologies based on their benefits and risks, as well as ethical and legal compliance. The European Commission-funded Coordination and Support Action ALIGNER (Artificial Intelligence Roadmap for Policing and Law Enforcement) will therefore identify promising AI technologies and propose a roadmap for future research investments in AI for LEAs. The stakeholders in ALIGNER are European actors concerned with AI, law enforcement and policing that collectively identify and discuss promising AI technologies for LEAs.

Although AI technologies provide many benefits for LEAs and the society in general, they also need to be trusted and accepted. Many organisations therefore propose recommendations to promote development of trustworthy AI that respect human rights and democratic values. For example, the High-Level Expert Group on Artificial Intelligence (AI HLEG), set up by the European Commission, identifies seven requirements for trustworthy AI: (1) Human agency and oversight, (2) Technical robustness and safety, (3) Privacy and data governance, (4) Transparency, (5) Diversity, non-discrimination and fairness, (6) Societal and environmental wellbeing, and (7) Accountability (European Commission, 2019).

This report describes the ALIGNER risk assessment of AI technologies for LEAs to identify and mitigate potential risks. The risk assessment is performed as a part of the ALIGNER AI Technology Watch method framework for impact assessment of AI technologies for LEAs (ALIGNER D3.1, Westman et al., 2022). The risk assessment complements the AI technology impact assessment (ALIGNER D3.1, Westman et al., 2022) and the fundamental rights impact assessment (ALIGNER D4.2, Casaburo & Marsh, 2023).

1.1 Gender Statement

ALIGNER partners actively safeguard gender equality and are aware of gender issues in science and technology (ref. "Commission of the European Communities: Women and Science: Excellence and Innovation–Gender Equality in Science, SEC (2005) 370, available at https://data.consilium.europa.eu/doc/document/ST-7322-2005-INIT/en/pdf).

ALIGNER monitors gender equality addressing biases and constraints throughout all the stages of the project as listed in Gendered Innovations 2 (ref "European Commission: Gendered Innovation 2 How Inclusive Analysis Contributes to Research and Innovation, (2020) available at https://op.europa.eu/en/publication-detail/-/publication/33b4c99f-2e66-11eb-b27b-01aa75ed71a1/language-en)

Outreach activities, visual representations, events, modes of data gathering and analysis and other research products related to D3.2 have been and will be gender proofed during the internal review process following the ALIGNER Gender policy (ref: ALIGNER D1.2 Project Handbook, section 8 'Gender aspects in publications and research').

1.2 Relation to Other Deliverables

ALIGNER offers a collaborative technology watch process, followed by three assessments: (1) the Technology Impact Assessment (ALIGNER D3.2 Westman, et al., 2022), (2) the Risk Assessment of AI Technologies, and (3) the Fundamental Rights Impact Assessment (ALIGNER D4.2 Casaburo et al., 2023). Both the Technology Impact Assessment method and the Fundamental Rights Impact Assessment have been successfully validated. This report will complement the existing documents where the complete integrated assessment approach will be used to assess the use of AI by LEAs for the benefit of society.

As well as dealing with the kind of inherent risks mentioned in this report, Law Enforcement Agencies must also consider and counter when necessary two other types of risk associated with AI technology. First, while the AI itself may comply with required principles, policy recommendations, laws and regulations, the purposes to which it is put by so-called ‘bad actors’ may create a crime or security threat. This can be done by using an AI to enable a crime to be carried out, either directly or indirectly. A second area of concern for LEAs could occur if opportunities arise for ‘bad actors’ to modify AI technologies to further their own ends where these may be criminal or security threats. Both of these will be considered in more detail in ALIGNER deliverable D3.3 – Taxonomy of AI Supported Crime.

1.3 Structure of this Report

This report begins with a review of relevant existing instruments for AI technology risk assessments in Section 2. The focus in this section is to assess risks associated with the development of AI technologies, but as some of the risk instruments include suggestions for responses to risks, mitigations are described when applicable. Section 3 complements Section 2 by putting greater focus on mitigations. The risk assessment and mitigations that are relevant for LEAs are then combined in the ALIGNER Risk Assessment Instrument presented in Section 4. Thereafter, Section 5 consists of examples of how the ALIGNER Risk Assessment Instrument can be used for a selection of risks and mitigation measures related to ALIGNER Scenario Cards.

1.4 Terminology

As this deliverable is focusing on risk assessment of AI technologies from LEA’s perspective, the following terminology has been implemented: ‘user’ refers to LEA and its personnel, ‘provider’ refers to the developer of the AI technology and ‘affected person’ refers to any individual who may be affected by LEAs using the system.

2. Instruments for AI Technology Risk Assessment (FOI)

AI systems are increasingly deployed for a wide range of applications that may have a large impact on individuals, organisations, and society. Policymakers, academics, standards bodies, industry, researchers, civil society organisations, and other stakeholders have therefore proposed different approaches for AI risk assessment (AIRA) (Ezeani et al., 2021). Each approach is developed specifically for the interests and objectives of concerned stakeholders. The AIRA enables stakeholders to have a proportionate approach that balances the benefits of AI with safeguards towards undesirable consequences of AI systems. Both Stahl et al. (2021) and Ezeani et al. (2021) review approaches for AIRA.

The wide scope of approaches for AIRA means that a complete review of all approaches is beyond the scope of ALIGNER. Further, AI systems, through data mining, are developed in phases of business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Shearer, 2000; Chapman et al., 2000; Brey et al., 2020a). Different approaches for AIRA are required for each phase. Given ALIGNER's objectives, this section only reviews approaches in the form of instruments for AIRA that are intended for public agencies in the deployment phase of AI systems.

This section summarises four instruments for AIRA:

- Assessment List for Trustworthy Artificial Intelligence (ALTAI) developed by the High-Level Expert Group on AI set up by the EU Commission,
- Shaping the ethical dimensions of smart information systems – A European perspective (SHERPA) developed as part of the SHERPA project, an EU Horizon 2020 project.
- Algorithmic impact assessment (AIA) developed by the Canadian government, and
- Artificial Intelligence Toolkit (AIT) developed by INTERPOL and the United Nations Interregional Crime and Justice Research Institute.

2.1 Assessment List for Trustworthy Artificial Intelligence (ALTAI)

The European Commission set up the High-Level Expert Group on Artificial Intelligence (AI HLEG) in 2018 (European Commission, 2019). The AI HLEG supported the Commission's vision of "ethical, secure and cutting-edge AI made in Europe" by publishing four influential deliverables to enable trustworthy AI (see ALIGNER D4.1, Eren et al., 2022). Trustworthy AI is important to build public confidence in AI. This maximises the benefits of AI-systems, whilst at the same time preventing and minimising their risks. Trustworthy AI is lawful, ethical, and robust from both a technical and societal perspective. In the view of the AI HLEG, trustworthy AI can be realised by fulfilling the following seven key requirements:

1. Human agency and oversight,
2. Technical robustness and safety,
3. Privacy and data governance,
4. Transparency,
5. Diversity, non-discrimination and fairness,
6. Societal and environmental wellbeing, and
7. Accountability.

Please see ALIGNER D4.1 for a comprehensive description of AI HLEG's seven key requirements for trustworthy AI (ALIGNER D4.1, Eren et al., 2022). Additionally, the AI HLEG operationalised the key requirements with the self-evaluation tool Assessment List of Trustworthy AI (ALTAI) (European Commission, 2020a). ALTAI helps organisations assess whether AI-systems that are developed, deployed, procured, or used, adhere to the key requirements. ALTAI is also available as an interactive online tool for self-evaluation that guides the user through each risk assessment and possible mitigations (European Commission, 2020b). The online tool is only a prototype that shows how ALTAI may be implemented in a structured and interactive format that guides the user through the assessment process. Although the online tool closely adheres to ALTAI, it also structures the questions so that follow-up questions are only shown when relevant, provides explanations as tooltips and provides overall assessments of risks and mitigations. Table 1 to Table 7 show examples of the ALTAI risks and mitigations for each of the seven key requirements. As the focus of this section is risks associated with AI, the cells concerning mitigations are coloured grey.

Table 1. Examples of risks and mitigations for human agency and oversight.

Requirement	**Risks**	**Mitigations**
Human agency and autonomy	Confusion whether interacting with human or AI system. Over reliance on AI systems. AI system risk creating human attachment, stimulating addictive behaviour or manipulating user behaviour. Human attachment to AI system. AI System deployed to manipulate and/or control user behaviour.	Reduce over-reliance. Reduce interference with decision-making. Avoid inadvertent effects.
Human oversight	Lack of training on how to exercise oversight.	Mechanism for detection and response for undesirable adverse effects. Procedure for safe abortion of operations. Determine how AI system is controlled or overseen and by whom.

Table 2. Examples of risks and mitigations for technical robustness and safety.

Requirement	**Risks**	**Mitigations**
Resilience to attacks and safety	Exposure to cyber-attacks (i.e., data poisoning, model evasion, or model inversion). Design or technical faults. Vulnerability towards attacks.	Certification for cybersecurity. Cybersecurity measures. Penetration testing. Security updates. Measures in place to ensure integrity, robustness, and security.
General safety	Damages from technical faults and misuse. Dependency on unreliable AI-supported decisions.	Risk management for use cases. Information about risks. Risk identification for technical faults and misuse. Definition of safety critical levels. Reliability testing. Fault tolerance. Safety review.
Accuracy	Adversarial consequences. Invalidation of data from operational use.	High quality data. Monitoring of accuracy. Information about accuracy.
Reliability, fall-back plans and reproducibility	Low reliability. Risks deriving from AI system using online continual learning. Irrelevant artefacts that skews performance.	Tests to ensure reproducibility. Verification and validation methods. Documentation (e.g. logging) to evaluate reliability and reproducibility. Fall back plans. Handling of low confidence scores by AI systems.

Table 3. Examples of risks and mitigations for privacy and data governance.

Requirement	Risks	Mitigations
Privacy	Right to privacy not respected	Reporting of privacy issues. Mechanisms that allow flagging issues.
Data governance	Use of or processing of personal data	Implementation of GDPR. Data protection impact assessment. Data protection officer. Oversight of data processing. Privacy by design. Data minimisation. Consent management. Life-cycle management. Alignment with standards.

Table 4. Examples of risks and mitigations for transparency.

Requirement	Risks	Mitigations
Traceability	Impossible to trace back which AI model led to a recommendation. Impossible to trace back which data was used by an AI system.	Assessment of input data quality. Assessment of output data quality. Logging practices.
Explainability	AI-driven decision impossible to understand by affected person.	Explanation of decisions. Survey understanding of decisions.
Communication	Inadequate communication to users.	Information about interactive system. Information about purpose of decision.

		Information about accuracy. Training and disclaimers.

Table 5. Examples of risks and mitigations for diversity, non-discrimination and fairness.

Requirement	**Risks**	**Mitigations**
Avoidance of unfair bias	Inclusion of inadvertent historic bias.	Strategies and procedures to avoid unfair bias. Diversity and representativeness in data. Tests for specific target groups. Tools for assessment of data quality. Monitoring of biases. Reporting of biases. Implement fairness.
Accessibility and universal design	Disproportional effects of outcomes. Unfairness.	Variety of preferences and abilities. Special needs or disabilities. Consider needs for assistive technologies.
Stakeholder participation	Stakeholders not included.	Inclusion of stakeholders.

Table 6. Examples of risks and mitigations for societal and environmental wellbeing.

Requirement	**Risks**	**Mitigations**
Environmental wellbeing	Negative environmental impacts.	Evaluation of environmental impact. Reduction of environmental impact.
Impact of work and skills	Impact on human work and work arrangements. De-skilling.	Inform impacted workers. Ensure understanding of impacts. Counteract de-skilling.

		Skill training.
Impact of society at large or democracy	Negative impact on society at large or democracy.	Assessment of societal impact. Minimize societal harm. Measures to ensure no negative impacts on democracy.

Table 7. Examples of risks and mitigations for accountability.

Requirement	**Risks**	**Mitigations**
Auditability	Inability to undergo audit.	Facilitate auditability. Third-party auditing.
Risk management	Ethical concerns. Conflicts of interest and values.	External guidance. Accountability measures. Risk training. Ethics review board. Identify value conflicts. Reporting of risks. Redress by design.

2.2 Shaping the Ethical Dimensions of Smart Information Systems – A European Perspective (SHERPA)

SHERPA was an EU Horizon 2020 project about the ethical and human rights implications of AI and big data (Brey et al., 2020b). The SHERPA guidelines for ethical use of AI and big data are intended to be actionable by organisations that use these systems. The requirements are directly based on the guidelines of the EU's AI HLEG (European Commission, 2019). However, minor adaptations were made to improve coherence and fitness for operationalization. Table 8 to Table 14 show the requirements, sub-requirements and examples of mitigations for each of the seven high-level requirements (as stated in the heading of each table).

Table 8. SHERPA requirements for human agency, liberty and dignity.

Requirement	**Sub-requirement**	**Examples of mitigations**
Human agency	Potential for impact on autonomy.	The system does not harm humans' autonomy. Use in decision-making is justified and minimised.

Negative liberty	Fundamental rights.	The system does not interfere with fundamental liberties.
Human dignity	Respect for human dignity.	The system does not affect human dignity negatively. The system is developed to promote human capacity.

Table 9. SHERPA requirements for technical robustness and safety.

Requirement	**Sub-requirement**	**Examples of mitigations**
Resilience to attack and security	Security, design, testing, and verification.	Evaluate and protect against security risks.
	Resilience.	Protection against successful attacks. Protection against substantial risks.
Fallback plan and general safety	Safety and verification.	Understanding of system functions and impact.
	Fallback.	Safety during system failures
Accuracy, reliability, and reproducibility	Accuracy, reliability, and effectiveness.	Ensure accuracy, reliability and effectiveness
	Reproducibility and follow-up.	Monitoring and documentation of security and safety objectives.

Table 10. SHERPA requirements for privacy and data governance.

Requirement	**Sub-requirement**	**Examples of mitigations**
Respect for privacy	Clarify roles and responsibilities towards information use, security and privacy.	Roles and responsibilities. Common culture that promotes ethical behaviour.
	Develop cultures of security and privacy awareness.	Culture of security and privacy awareness. Log of information access.
	Personal data use, reduction, and elimination.	Minimise use of personal data. Changes from sensitive to non-sensitive data.
	Personal data storage.	Protection of data storage according to sensitivity.
	Informed consent.	Collection of personal data with informed consent or by other legal means.
	Creation of new personal data.	Protection of new personal data.
	Subsequent collection and/or creation of new personal data.	Collection of new personal data only when necessary.
	Privacy awareness.	Users can flag privacy issues.

		Notice and control over personal data.
	Data review and minimization.	Oversight mechanism for data storage. Protection of personal data.
	Alignment with existing standards.	Alignment with standards for data management and governance.
	Data Protection Officers.	Involve Data Protection Officer.
Quality and integrity of data	Oversight of data quality.	Quality and integrity of personal data. Governance and management of data assets.
	Employment of protocols and procedures for data governance.	Protocols for data governance. Safeguards for compliance with protocols.
Access to data	Oversight of access to data.	Qualifications for data access. Log of data access.
	Availability of data.	Process to remove and correct data.
	Protection against re-identification.	Protection against de-anonymization.
Data rights and ownership	Clarity on ownership of data.	Clarification of ownership of personal data.

Table 11. SHERPA requirements for transparency.

Requirement	**Sub-requirement**	**Examples of mitigations**
Traceability	Traceability measures.	Traceability of development.
	Responsibility for traceability.	Human intervention to prevent harmful outcomes.
Explainability	Training data.	Compliance with ethical standards.
	Explainable systems.	Decision transparency.
	Explanations of rationale.	Rational for system choices. Reasons for collection and use of personal data.
	Trade-offs.	Trade-off between explainability and performance.
Communication	Communication regarding interactions with systems.	Information that algorithmic system makes decisions.
	Communication with stakeholders.	Open communication with stakeholders. Information about system capabilities. Purpose and benefit of system. Understandable usage scenarios.

	Communication within end-user and stakeholder community.	Culture of mutual trust.

Table 12. SHERPA requirements for diversity, non-discrimination, and fairness.

Requirement	**Sub-requirement**	**Examples of mitigations**
Avoidance and reduction of harmful bias	System bias assessment.	Data representativeness. Evaluation of system biases.
	Use bias assessment.	Avoid creating biases.
	Engagement with users to identify harmful bias.	Users can flag biases. How the system may affect individuals. Methods for redress.
	Anticipating harmful functional bias.	Avoid non-intended use cases.
	Decision variability.	Impact of decision variability on fundamental rights.
	Avoiding harmful automation bias.	Meaningful human control. Prevention of overreliance.
Ensuring fairness and avoidance of discrimination	Accessibility and usability.	Accessibility for users with assistive technologies.
	Intended use.	Reasonable function of algorithm.
	Review process.	Risk assessment.
	Distributing the system to end-users.	Information about accuracy and errors.
	Whistleblowing.	Process to anonymously inform external parties.
Inclusionary stakeholder engagement	Diversity.	Participation of stakeholders.
	Inclusion.	Inclusion of diverse viewpoints.

Table 13. SHERPA requirements for individual, societal, and environmental wellbeing.

Requirement	**Sub-requirement**	**Examples of mitigations**
Sustainable and environmentally friendly systems	Environmental impact.	Evaluation of ecological impact.
Individual wellbeing	Individual wellbeing assessment.	Assess impact on individual wellbeing.
	Emotional attachment.	Minimize unwanted attachment.
Societal wellbeing	Societal impact assessment.	Assess impact on social relations.

	Engagement with stakeholder community.	Evaluation of social impact. Understanding of social impact.
Democracy and strong institutions	Mitigation of impacts on democracy.	Assess impact on political processes.

Table 14. SHERPA requirements for accountability.

Requirement	**Sub-requirement**	**Examples of mitigations**
Auditability	Engagement and reporting.	Incident reporting. Proactive problem management. Culture of risk awareness. Performance indicators.
	Compliance as culture.	Culture of compliance awareness. Facilitation of auditability.
	Code of ethics.	Culture of internal auditing.
Minimising and reporting negative impacts	Reporting Impact.	Risk assessment. Promote accountability.
	Minimising negative impact.	Balance of risks and benefits.
Internal and external governance frameworks	Impact on business.	Impact on decision-making processes. Rationale for using the system.
	Identify interests and values at risk.	Trade-offs between interests and values.
Redress	Redress mechanisms.	Process for redress.
Human oversight	Avoiding automation bias.	Level of human control. Prevention of overreliance.
	Responsibility.	Identification of human control. Human control for protection.

2.3 Algorithmic Impact Assessment (AIA)

The Algorithmic Impact Assessment (AIA) tool was developed by the Canadian government to support its directive on automated decision-making (Secretariat, 2021). The tool is based on a framework by the AI Now Institute for algorithmic impact assessment to ensure public agency accountability (Reisman, 2018). This framework respects the public's right to know about AI systems that impact their lives, increases public agencies' ability to evaluate AI systems, ensures accountability by external review and ensures an opportunity to dispute the use of AI systems (Government of Canada, 2023a).

The AIA tool is a web-based questionnaire available online (Secretariat, 2023), as well as a standalone system (Government of Canada, 2023b). The AIA is organised according to policy, ethical, and administrative law considerations of the Government of Canada applied to the context of automated decision-making. The AIA is based on consultations between the Treasury Board of Canada Secretariat with public institutions, academia, and civil society. The AIA is designed to help departments and agencies to better understand and manage risks associated with automated decision systems. Its questionnaire consists of about 80 risks and mitigation questions. It calculates an impact score based on the level of risk for each question and suggests mitigations for how the risks are managed. The mitigation score must reach more than 80 % of the maximum score to reduce the impact score. The impact is classified in levels from no impact to very high impact based on the impact score. The impact levels determine the mitigations required under the Directive on Automated Decision-Making according to the AIA tool (Secretariat, 2021).

Table 15 shows the risk areas and their description. The AIA also assesses automated decisions on a broad range of topics. Table 16 shows the mitigation areas, their description, and topics in the AIA.

Table 15. Risk areas for AIA.

Risk area	**Description**		**Topics**
Project	Project phase	Project owner, description, and development stage.	Point of contact. Project phase. Project description.
	Business drivers	Reasons for introducing automation into the decision-making process.	Motivation of automation. Client needs. Public benefits. System effectiveness. System benefits. Confinement to only client needs. Trade-offs between client interests and program objectives. Alternative non-automated processes.
	Risk profile	High-level risk indicators for the project (e.g., vulnerability of clients).	Area of public scrutiny. Vulnerable clients. High stake decisions.

			Impact on number of personnel or their roles. Barriers for persons with disabilities.
	Project authority	Need to seek new policy authority for the project.	Policy authority.
System	About the system	Capabilities of the system.	Image and object recognition. Text and speech analysis. Risk assessment. Content generation. Process optimization and workflow automation.
Algorithm	About the algorithm	Limitations on disclosure of the algorithm. Ability to explain how it arrives at outputs.	Algorithm characteristics. Algorithm is a trade secret. Interpretability and explainability of the algorithm.
Decision	About the decision	Classification and description of the decision being automated (e.g., health services, social assistance, licensing).	Decisions that will be automated. Decisions regarding health, economic interests, social assistance, access and mobility, licensing and permits, employment.
Impact	Impact assessment	Type of automation (full or partial). Duration and reversibility of the decision. Areas impacted (e.g., rights, privacy and autonomy, health, economic interests, the environment).	Type of automation. Role in decision-making process. Decisions based on judgement or discretion. Evaluation criteria. System output and interpretation. Replacement of human assessment. Reversibility of decisions. Duration of impacts.

			Impacts on individual rights, freedoms, health, wellbeing, economic interests, sustainability.
Data	Source	Provenance, method of collection, and security classification of data used by the system.	Use of personal information. Security classification. Control of data. Number of sources. Use of data from the Internet or other IT-systems. Data collection for training. Data collection for system input.
	Type	Nature of the data used as structured or unstructured (audio, text, image, or video).	Analysis of unstructured data. Types of unstructured data.

Table 16. Mitigation areas for AIA.

Mitigation area	**Description**		**Topics**
Consultations	Internal and external stakeholders	Internal and external stakeholders consulted. Digital policy teams. Subject matter experts in other sectors.	What groups. Which stakeholders.
De-risking and mitigation measures	Data quality	Processes to ensure that data is representative and unbiased, as well as transparency measures related to those processes.	Tests against biases. Documentation for resolving data quality issues. Gender bases analysis. Accountability for system. Risk management of unreliable data.
	Procedural fairness	Procedures to audit the system and its decisions,	Audit trail.

		as well as the recourse process.	Links between decisions and legislation. Change log of model and system. Reasons for decisions. System access. Capture user feedback. Process to challenge decisions. Human override of decisions.
	Privacy	Measures to safeguard personal information used or generated by the system.	Privacy Impact Assessment. Security by design. No connections to other systems.

2.4 Artificial Intelligence Toolkit (AIT)

The Artificial Intelligence Toolkit (AIT) was created by INTERPOL and the United Nations Interregional Crime and Justice Research Institute (UNICRI). The toolkit aims to help LEAs address the most pressing challenges when it comes to the use of AI. The addressed challenges are based on the need for guidance. The envisioned primary users of AIT are personnel of law enforcement agencies at all levels (UNICRI and INTERPOL, 2023a). The Risk Assessment process in AIT includes five steps: preparing, assessing, interpreting, communication and maintaining.

The AIT consists of a user´s guide and seven individual resources where the supporting document “the Principles for Responsible AI Innovation” (UNICRI and INTERPOL, 2023c) is the foundation for the entire AI Toolkit and guides LEAs in incorporating AI systems into their work with good AI ethics, policing practices and respect for human rights. The five core principles for Responsible AI Innovation for law enforcement community used in the AIT are:

1) Lawfulness (LEAs must follow the applicable laws and regulations throughout the design, development, and use of AI systems),
2) Minimization of Harm (LEAs prevent, eliminate, or mitigate the risk of harm to individuals and communities that can arise in the context of AI development, procurement and use),
3) Human Autonomy (LEAs engage with AI in a way that safeguards humans’ capacity and right to self-governance),
4) Fairness (LEAs should ensure, throughout their engagement with AI systems, a just and non-discriminatory treatment of individuals and groups and should contribute to a more equitable society), and

5) Good Governance (means that agencies should aim to set up an overarching structure for audits and accountability and to foster a culture of responsible AI innovation).

"The Responsible AI Innovation in Action Workbook" (UNICRI and INTERPOL, 2023d) intends to support LEAs through the whole AI life cycle that generally includes three main stages: (1) planning; (2) development / procurement; (3) use & monitoring. It contains structured exercises to help agencies along the path towards responsible AI innovation. It also contains a questionnaire, "The Risk Assessment questionnaire", that intends to support LEAs to estimate the risks an AI system may pose from a responsible AI innovation perspective. More specifically, it supports LEAs with identifying the potential adverse impacts on society as a whole, groups and individuals, as well as the probability of such impacts occurring.

UNICRI and INTERPOL(2023c) has been used to define the risks and mitigations in Table 17. As the focus for this section are risks associated with AI, the cells for mitigations are coloured grey.

Table 17. Examples of risks for lawfulness (LEAs must follow the applicable laws and regulations throughout the design, development and use of AI systems),

Requirement	**Risks**	**Mitigations**
Legitimacy	LEA interfere with people's rights without a valid reason based on law and standards.	Ensure a legal basis for interference. Ensure following a legitimate goal.
Necessity	LEA interfere with people's rights when not needed to fulfil the identified legitimate goal.	Ensure that the legitimate goal cannot be achieved without interfering with human rights.
Proportionality	LEA interference not proportionate.	Proportionality assessment.

Table 18. Examples of risks for minimization of harm (LEAs prevent, eliminate or mitigate the risk of harm to individuals and communities that can arise in the context of AI development, procurement and use),

Requirement	**Risks**	**Mitigations**
Robustness and Safety	Unreliable system. Unsecure system. Unsafe system.	Ensure AI systems can perform intended function adequately and cope with changes in its environment. Ensure protection against attacks. Safeguards to prevent unacceptable harm and minimize unintentional and unexpected harm.
Accuracy	Incorrect predictions, recommendations or decisions.	Verification of accuracy.

		Training of AI system with sufficient and good quality data. Training of user. Mindful of the origin and composition of the training data. Testing the system by independent third parties.
Human and environmental wellbeing	LEA not preserving and improving the welfare of people and the environment in their AI innovation journey.	Examination of direct and indirect consequences.
Efficiency	Costs overweight the benefits of using a certain AI system.	Needs and capabilities assessment.

Table 19. Examples of risks for human autonomy (LEAs engage with AI in a way that safeguards humans' capacity and right to self-governance),

Requirement	**Risks**	**Mitigations**
Human control and oversight	Lack of human control and oversight.	Verify that the AI systems are built with functionalities that ensure that humans remain in charge during use. Certify that humans have the last word regarding certain decision.
Human agency	Over reliance on AI systems. Limited access to information and/or opportunities. AI system deployed to manipulate and/or control user behaviour. AI system malfunction result in lacking information for humans in decision making.	Certify that the AI systems do not compromise the ability of the users of those systems to act and make decisions independently. Training how properly engage. Frequent check-ups.
Privacy	Interference with the right to privacy.	Protect and limit interference in the private sphere of individuals.

		Use privacy-by-design and privacy-enhancing technology.
Transparency and Explainability	Lack of awareness and insight of the AI system, its use and consequences.	Verify that the providers of AI system disclose all the necessary information and documentation to its users. Promoting good communication practices.

Table 20. Examples of risks for fairness (LEAs should ensure, throughout their engagement with AI systems, a just and non-discriminatory treatment of individuals and groups and a contribution to a more equitable society)

Requirement	**Risks**	**Mitigations**
Equality and non-discrimination	Discrimination and inequality. Wrongful and unjustified outcomes.	Ensure equal treatment and opportunities for all. Refrain from unjustifiably discriminating. Ensure appropriate quality and quantity of training data.
Protection of vulnerable groups	Disadvantage or disproportionately negative impact and harm to certain groups.	Safeguards in place for protection. Equal access and opportunities or benefits for vulnerable groups.
Diversity and Accessibility	AI system accessible only by a narrow range of groups and individuals.	Building inclusive systems designed in a user-centric way.
Contestability and Redress	Affected persons cannot challenge decisions. Impossible to argue against AI-supported decisions.	Necessary technological and organizational measures are in place.

Table 21. Examples of risks for good governance (means that agencies should aim to set up an overarching structure for audits and accountability and to foster a culture of responsible AI innovation).

Requirement	**Risks**	**Mitigations**
Traceability and Audibility	Impossible to prevent, identify or resolve negative consequences that might arise from AI use.	Set up requirements, procedures and technical solutions to ensure traceable decision-making processes.

	Impossible to supervise the development and use of AI system. Decisions not traceable.	Adequate documenting of decisions that influence AI systems’ outputs. Tracking and documenting AI outputs, including the input data used, the model and parameters selected, the model’s output, the user’s name, date and any other relevant information. Ensure essential elements can be assessed by internal or external auditors.
Accountability	Responsible persons not identified.	Mechanisms and processes to enable determination of responsibility and accountability.

3. AI Technology Risk Assessment and Mitigations (CBRNE)

Risks associated with AI technology encompass a spectrum of concerns. As previously shown in Section 2, it is possible to identify a number of different risks and categorise them in several ways. Several challenges may occur. AI systems can inherit biases present in training data, leading to discriminatory outcomes. Ethical considerations mandate the assessment of these biases to ensure fairness and prevent undue harm. Security risks are another dimension, with the potential for malicious actors to exploit vulnerabilities in AI systems for their gain. Furthermore, there are concerns related to accountability, transparency, and unintended behaviour, requiring in-depth evaluation to ensure responsible AI deployment (Ulnicane, 2022).

Risk assessment and mitigation in the context of AI involve a systematic process of identifying, analysing, and addressing potential threats, vulnerabilities and adverse outcomes that may arise from the development, deployment, and use of AI technology. This multifaceted approach aims to minimize or eliminate the negative consequences of AI while maximizing its benefits (European Commission, 2020c).

The risk associated with AI technology can be categorised according to various aspects and areas. Although some categories might naturally overlap or complement each other, in this section the risks and mitigations of AI technologies are categorised as follows (Burgess & Kloza, 2021):

- Lawfulness, fairness and transparency of processing
- Data and storage minimisation
- Data accuracy and security
- Data subject rights and access control
- Automated decision-making

It is important to note that the risk assessment of AI technologies in relation to fundamental rights, freedoms and ethical considerations has been covered in some detail in ALIGNER deliverable D4.2 (Casaburo & Marsh, 2023), therefore this category has not been covered in this document for the sake of avoiding repetitions and overlaps.

Mitigating AI risks encompasses a range of proactive measures aimed at minimizing the identified vulnerabilities and potential threats. From robust data privacy and security protocols to the implementation of fairness-aware algorithms and explainable AI techniques, various strategies need to be employed to enhance the reliability, multifaceted transparency, and safety of AI systems (Glauner, 2022). Furthermore, continuously updating and monitoring AI models, incorporating human oversight in critical decision-making, and fostering collaboration between AI providers, domain experts, and LEAs are critical aspects of effective risk mitigation (Mueck et al, 2023). Policymakers, law enforcement agencies, technology providers, and civil society must work together to ensure that AI technologies are integrated responsibly and ethically in policing and law enforcement.

This section outlines in a series of tables examples of risks associated with implementation of AI solutions as identified in a literature review. Furthermore, the corresponding technical and operational mitigation measures are identified and presented for each risk. This section focuses on mitigation measures, but to put these into context, risks are also exemplified. Within each table, the cells containing the risks are coloured grey.

As this deliverable is focusing on risk assessment of AI technologies from LEA's perspective, the terminology presented in Section 1.4 has been implemented. The term 'applicable law' in this section refers to the body of EU regulations and directives, plus any other national or international applicable law in the relevant country.

The risk and mitigations presented in the subsections following below have been crystallised from an extensive literature review conducted across the various publications in the field of AI, trustworthiness, and transparency. Each of the identified risks and associated mitigation strategies have been consolidated as a result of the respective authors' view presented through the sources.

3.1 R1 – Lawfulness, Fairness and Transparency of Processing

Lawfulness, fairness, and transparency of data processing are key principles that guide the responsible, legal and ethical use of AI systems in the context of policing and law enforcement. They ensure that AI processing complies with applicable laws, promotes fairness in decision-making and maintains the transparency of its operations. Each of these principles are further examined below in the context of law enforcement (Truby et al, 2022):

- Lawfulness: AI systems must operate within the bounds of applicable laws governing fundamental rights, data protection, privacy, and other relevant areas. For instance, this principle requires LEAs to collect, store and process personal data only when necessary for the performance of a task carried out in the public interest for law enforcement purposes. Adhering to legal requirements helps ensure that AI systems are used in a manner that respects individual fundamental rights and safeguards sensitive information.
- Fairness: Fairness in AI refers to ensuring that the outcomes and decisions produced by AI systems do not systematically disadvantage or discriminate against individuals or groups. Bias and discrimination can arise due to various factors, such as use of biased training data, flawed algorithms, or unfair decision-making processes. Providers and LEAs should strive to identify and mitigate biases to promote fair and equitable treatment of individuals. This can involve techniques like data pre-processing, algorithmic auditing, and validation to reduce bias and prevent unfair outcomes (Jacobs & Simon, 2022).
- Transparency: Transparency in AI involves making the decision-making process and the factors influencing AI outcomes in law enforcement and policing clear and understandable to individuals and stakeholders. It includes providing explanations for AI decisions and ensuring that the rationale behind those decisions can be effectively understood and communicated by LEAs. Transparent AI systems enable individuals to understand how their data is processed and how decisions that affect them are made, so as to allow an effective exercise of due process and rights in court. Enhanced transparency helps build trust, promotes accountability, and allows for effective scrutiny of AI systems.

By upholding the principles of lawfulness, fairness, and transparency, both providers and LEAs can foster trust in AI systems, mitigate risks and ensure that AI technologies are developed and used responsibly and ethically (European Commission, 2020a). The following table outlines the risks and technical and operational measures to ensure lawfulness, fairness, and transparency of AI systems in the context of law enforcement and policing.

Table 22. Risks and mitigations measures for lawfulness, fairness, and transparency of processing in AI systems.

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
R1.1	**Fairness and transparency** Violation of the pillars of (i) fairness (the ability to treat data subjects in a manner that respects their human rights and treats them equally without undue biases), (ii) accountability (the ability to explain the system's decision-making and reasoning processes both in general and regarding specific outcomes) and (iii) transparency (the ability to disclose details on the system processes).	The user needs to verify that adequate measures have been taken on the provider end to ensure that the design process was centred around these pillars from an early stage, that this process has been documented and that results of this process are evidenced and build the foundations of the system.	Before deployment, the system should be inspected by independent experts for scrutiny. Reporting structures should be in place to verify integrity and handle such concern regardless of the origin (i.e., where, how, or by whom it has been raised).	(Estella, 2023) (Burgess & Kloza, 2021) (Jacobs, 2022)
R1.2	**Transparency in the logic involved** Lack of transparency in the system's reasoning.	The user needs to ensure that the provider has documented and clearly elucidated all automated processes in an accessible way so that the algorithmic logic is clear and understandable.	Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment.	(Burgess & Kloza, 2021) (Estella, 2023)
R1.3	**Transparency in documentation** Lack of clarity in documentation of operational procedures.	Implementation of algorithms should adopt transparency and well documented instructions to enable subsequent forensic audits.	Users are responsible to engage in a training activity to enable technology transfer workshop prior to the deployment of the system.	(Burgess & Kloza, 2021) (Estella, 2023) (European Commission, 2020a)
R1.4	**Explanation of procedure**	Algorithms should include enough details on the internal	User is to be trained to understand the processes and their	(Lorch et al, 2022)

	Faulty or lacking analysis of logs/procedures and system processes that may be requested.	operations to facilitate subsequent data audits. The user needs to ensure that the provider has clearly documented all procedures and processes involved in the operation of AI systems and standardized the explanation procedures to ensure consistency and reduce the risk of errors due to miscommunication or misunderstanding.	impact on investigations and operational aspects.	(European Commission, 2020c) (Burgess & Kloza, 2021)
R1.5	**Auditing and ex-post inspection** Failure to make information on the system programming and functioning available on request for inspection by skilled professionals.	The user needs to ensure that the provider has implemented forensic logs with enough details to carry out system wide audits.	Users are responsible to check the system logs during the testing phase within an operational environment.	(Burgess & Kloza, 2021) (Truby et al, 2022) (Krakovna et al, 2020)
R1.6	**Early warning system** Failure to regularly review and assess components in view of their accuracy, fairness and to detect potential risk for unintended outcomes.	The user needs to ensure that the provider has integrated a periodic review of the models used in the computational process of information. The time bound review for improvements should be notified to the technology provider.	User should perform periodic tests to ensure the robustness and reliability of the system.	(Lorch et al, 2022) (Burgess & Kloza, 2021) (Hupont et al, 2023)
R1.7	**Non-defined purpose or incompatible further processing** The collection and processing of personal data is not in agreement with the purpose and legal ground.	The user needs to ensure that the provider sufficiently restricts the processing of data in relation to defined purposes. The collection and retaining of the data should be kept to a minimum and	The nominated data security governing body is to ensure that use of data is limited in accordance with the strictest standards possible. Technical experts are responsible for implementing these	(Truby et al, 2022) (Estella, 2023) (Hupont et al, 2023)

	Database integration and additional restrictions might not have been provided to keep data that was collected exclusively, e.g., for counter-terrorism measures, to not be accessible by general user.	only in relation to the intended purpose.	procedures at architecture and infrastructure level, and to document these in a transparent way. User is to be trained to both, comply with set policy and flag any possible or suspected breach.	

3.2 R2 – Data and Storage Minimisation

Data lies at the heart of AI systems, enabling them to learn, adapt and make informed decisions in policing. However, the sheer volume of data collected and stored raises concerns about privacy breaches, security vulnerabilities and the potential for misuse. Data and storage minimization entails collecting, processing, and storing only the necessary data required to achieve the AI system's intended goals. By reducing the data footprint, the risks of unauthorized access, breaches and unintended processing are mitigated (Laux, 2023a).

Risk assessment and mitigation for AI in relation to data and storage minimization involves evaluating and addressing potential risks associated with the collection, storage, and use of data by AI systems. Data and storage minimisation aligns closely with ethical principles and laws surrounding data protection, such as the General Data Protection Regulation (GDPR) and the Law Enforcement Directive (LED) in the European Union. Minimisation not only safeguards individual privacy but also enhances the overall transparency and trustworthiness of AI systems. The aim of data storage minimisation is to minimise the amount of data collected and stored while ensuring that the data used is relevant, accurate and secure (European Commission, 2020c). Mitigation measures outlined in the following table aim to help eliminating risks such as data breaches, privacy violations and potential misuse of sensitive information.

Table 23. Risks and mitigation measures for data and storage minimisation in AI systems.

#	***Risk description***	***Risk mitigation measure suggested - Technical***	***Risk mitigation measure suggested - Operational***	***Sources***
R2.1	**Data minimization not applied** The collection and processing of personal data is not adequate, relevant, and limited (disproportionate) to what is necessary in relation to the purposes	Algorithms should be validated for the minimum amount of information required to successfully demonstrate the security functions for which the algorithms were originally	Regular reviews (e.g., monthly) of the system/procedure should be implemented to ensure that the type and amount of data collected is the absolute minimum of what is required to perform tasks at hand.	(Bostrom & Yudkowsky, 2022) (Jacobs, 2022)

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
	for which they are processed and as described in the public privacy policy. Access control mechanisms do not adequately limit user access to only the necessary data, i.e., data needed to accomplish their tasks.	intended. The user needs to ensure that the system providers implement efficient access control mechanisms allowing authorised users to configure access rights based on the type of data, the profile of the system user and the purpose of the task, and that safeguards are in place that prevent as well as flag any breach, whether accidental or purposeful.	All users involved should be adequately trained to be able to comply with applicable laws and to notice and flag when data is unnecessarily collected.	(Laux et al, 2023a)
R2.2	**Broader scope** For systems that request information only on a "per case" basis (through either a human operator or relying on automated processing): Failure to limit access to only the necessary part of metadata assigned to that case	Algorithms requiring access to historical records should be limited to the specific cases that are under consideration.	Users should validate the system against cross-information access among different cases.	(Laux et al, 2023b) (Burgess & Kloza, 2021) (Hadzovic et al, 2023)
R2.3	**Unauthorized disclosure of sensitive data** Data with sensitive qualities makes individuals identifiable to unauthorised figures.	The combination of purpose-designed system architecture, tailored algorithms, encryption methods and access control measures should be implemented to prevent such incident.	All users with access rights should be aware of and comply with their data safeguarding duties. Alert mechanisms should be put in place, on both human and technological level, to flag and identify any possible weak points or unauthorised access attempts. Environmental factors (e.g., specific background, presence of	(Truby et al, 2022) (European Commission, 2020c) (Burgess & Kloza, 2021)

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
			unauthorised humans, etc) that could jeopardise prevention of this risk should be identified and eliminated.	
R2.4	**Redundant data** Failure of deleting redundant data, and/or after a predefined timeframe either automatically or by presenting a reminder or prompt.	Mechanisms on technology level, should be put in place to "double up" checks and alert prompts, and to flag/forward these in case of no response.	Mechanisms on human level, should be put in place to "double up" checks and alert prompts, and to flag/forward these in case of no response.	(Burgess & Kloza, 2021) (Yampolskiy , 2020) (Truby et al, 2022)
R2.5	**Anonymization and pseudonymisation** Failure to apply anonymization and pseudonymisation techniques in time.	Architecture protocol should establish ways of detecting such failure ensuring that remedial contingency measure is put into place promptly upon discovery.	User protocol should establish ways of detecting such failure ensuring that remedial contingency measure is put into place promptly upon discovery.	(Lorch et al, 2022) (Laux et al, 2023b) (Yampolskiy , 2020)

3.3 R3 – Data Accuracy and Security

The quality, accuracy and integrity of data directly influence the outcomes generated by AI algorithms. However, data is vulnerable to errors, manipulation, and breaches, raising concerns about the reliability of AI-driven decisions. Ensuring that data is accurate and secure is critical to harness the potential benefits of AI in policing while mitigating potential harms (Yampolskiy, 2020).

Inaccurate, noisy, or biased data can lead to following unwanted or potentially harmful outcomes (Sanz-Urquijo et al., 2022), (Estella, 2023):

- Perpetuating discrimination
- Incorrect predictions or actions
- Failure to adapt to new patterns, resulting in degraded performance
- Decreased reliability of AI-driven decisions
- Introducing vulnerabilities that can be exploited by malicious actors
- Non-compliance with data protection regulations or legal standards
- Violating ethical principles and creating negative consequences for individuals or society

AI risk assessment in relation to data accuracy involves evaluating and mitigating the potential risks that arise from inaccuracies, errors, biases, or inconsistencies in the data used to train and operate AI systems. Ensuring data accuracy through comprehensive data cleaning, pre-processing and bias detection techniques is essential to mitigate these risks and promote fairness (Hupont, 2023).

Data security is another critical aspect of AI for policing, as compromised data can lead to following consequences:

- Unauthorized access to sensitive data or breaches of the data storage infrastructure
- Privacy violations
- Incorrect decision-making
- Data tampering or manipulation

A comprehensive approach to AI risk assessment involves identifying vulnerabilities related to data security, implementing robust mitigation security measures, and continuously monitoring and updating security protocols to adapt to evolving threats.

The following table outlines identified risks and mitigation measures in relation to the data accuracy and security in AI systems.

Table 24. Risks and mitigations measures for data accuracy and security in AI systems.

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
R3.1	**Data governance** Insufficiently established data governance.	Implementation of regulatory compliance for the use of technology in collecting, processing and interpreting data.	Technology/system should be embedded in structures that allow the nominated data security governing body to enforce established policy as part of the overall data management strategy. An independent body should be put in place to regularly and critically review the quality of governance in the context of applicable law.	(Yampolskiy, 2020) (Sanz-Urquijo et al, 2022)
R3.2	**Data update** Failure of automatic data update mechanism.	System architecture should be designed so that warning is issued as soon as automatic updates are missed or were disabled.	Regular "manual" checks for existence of updates should be performed.	(Burgess & Kloza, 2021) (Hupont et al, 2023)

#	Risk description	Risk mitigation measure suggested - Technical	Risk mitigation measure suggested - Operational	Sources
				(Glauner, 2022)
R3.3	**Inaccurate data** Flaw in technical procedures that enable reviewing and removing inaccurate or outdated data.	Algorithms should validate the input data format against corruption prior to processing the information.	Implementation of robust data quality assurance procedures that involve continuous monitoring, validation, and verification of the data used by the AI system. Data quality checks should be conducted at various stages, from data collection to model deployment.	(Yampolskiy, 2020) (Burgess & Kloza, 2021) (Laux et al, (2023a)
R3.4	**Data tagging** Error in tagging and marking procedures designed to allow for the marking of different files and datasets to illustrate their reliability, origin, file type, sensitivity, and usage rights. This, consequently, could prevent clarity of limits to data or could impair quality of information on the data to be processed	The algorithm models should be validated for the robustness of system implementation.	The data set used in the training of algorithms should be audited to ensure the correctness.	(Yampolskiy, 2020) (Hupont et al, 2023) (Lorch et al, 2022)
R3.5	**Data encryption** Error or failure on a level of data encryption and/or other privacy enhancing technologies (PETS)	The internal data stored by the algorithm and the external data accessed by the algorithms such as models, should be encrypted and protected against external changes and parameter manipulation.	Establish regular review for checking functionality, ensuring coherence, and updating of encryption method.	(Burgess & Kloza, 2021) (European Commission 2020c)

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
				(Laux et al, 2023a)
R3.6	Database system's failure to appropriately encrypt stored information as well as integrated automated backups by default.	Robust and widely recognized encryption algorithms, such as AES (Advanced Encryption Standard) for symmetric encryption and RSA (Rivest–Shamir–Adleman) for asymmetric encryption should be integrated.	Establish regular review for checking functionality, ensuring coherence, and updating of encryption method.	(Burgess & Kloza, 2021) (Estella, 2023) (Laux et al, 2023a)
R3.7	Insufficiently high security and system safety standards and resulting failure to guarantee the ongoing confidentiality, integrity, availability and resilience of processing systems and services.	Transport Layer Security (TLS) protocols should be used for securing data in transit over networks. TLS ensures that data exchanged between systems is encrypted and protected from eavesdropping.	Users should ensure that sufficiently high security and system safety standards that comply with current law and recommendations are put in place and are updated regularly.	(Yampolskiy, 2020) (European Commission, 2020c)
R3.8	Insufficient action in the event of a data breach.	Appropriate protocols should be implemented to capture necessary details on the event that can be retrieved and exported to allow for further investigation or to inform supervisory authorities and affected persons of possible consequences.	It is of utmost importance that protocol for the event of data breach is established by the respective authority and that infrastructure and resources to follow protocol are put in place and reviewed.	(Truby et al, 2022) (Burgess & Kloza, 2021) (Krakovna et al, 2020
R3.9	**Classification of data subjects** Fault in labelling system that assigns different types of data categories to any data	Algorithms should be designed so that the AI system is not influenced by historical records, e.g. of crimes committed.	Establish regular review for ensuring coherence and unbiased nature of the AI systems.	(Burgess & Kloza, 2021)

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
	related to individuals, e.g., based on their involvement in a crime or their previous interactions with the justice system.			(Hupont et al, 2023) (Laux et al, 2023a)
R3.10	**Inferences** For systems that allow users to mark data as factual or data based on personal assessments: failure to properly label or process so as to enable the distinction of different types of data given their quality.	A user centred approach in the system design process should ensure that user-friendly interfaces make it easy to handle data appropriately.	User should be provided with training and regular updating of skills needed to be familiar with the system, their duties, and relevant applicable law/policy.	(Truby et al, 2022) (Sanz-Urquijo et al, 2022) (Glauner, 2022)
R3.11	**Special categories of data** Flagging error in system datasets that contain special categories of data (that is, data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data used for identification purposes, health data and/or data concerning sexual orientation).	Differential privacy techniques and anonymisation solution should be applied to anonymise or pseudonymization of the special categories of data to protect individuals' privacy while still allowing meaningful analysis of aggregated data.	Possible sources of such flagging error should be established, and mitigation/contingency should be defined for each source as part of data governance. Examples for sources are human error, algorithm error, system intrusion, recognition error, etc.)	(Burgess & Kloza, 2021) (Hupont et al, 2023) (Glauner, 2022)
R3.12	**Privacy by design** Privacy considerations are not adequately integrated into the	System sign on authentication and authorisation activities to be carried out and ensure the system does not provide any back-door	It is a legal obligation that the Data Protection Impact Assessments (DPIAs) for AI technologies in the context of LEAs and policing must be	(Burgess & Kloza, 2021)

#	Risk description	Risk mitigation measure suggested - Technical	Risk mitigation measure suggested - Operational	Sources
	design and architecture from the outset.	entry to unauthorised users. Users are to verify that adequate measures have been taken on the provider end to include the listed concerns in their system design approach at an early stage, that this process has been documented and that results of this process are evidenced, and reflected in the system. Before deployment, the system should be inspected by independent experts for scrutiny.	conducted to identify potential privacy risks and implement measures to address them before deployment. Protocol should be established for cases that flaws in the design process are detected.	(Estella, 2023) (European Commission (2020c)

3.4 R4 – Data Subject Rights and Access Control

AI systems often process vast amounts of personal data, making it essential to conduct a thorough risk assessment to protect data subject rights and ensure proper access control. Users of the AI systems are responsible to safeguard the rights of individuals whose data is being utilized. Central to this concern is the concept of data subject rights – the rights individuals have over their personal data (Hupont et al, 2023).

Data subject rights encompass a range of rights granted to individuals under data protection laws, such as the European Union's General Data Protection Regulation (GDPR) and the Law Enforcement Directive (LED), and similar laws worldwide. These rights include the right to access personal data, rectify inaccuracies, erase data ("right to be forgotten"), object to processing, restrict processing and data portability. Data subject rights empower individuals to maintain control over their personal data and how it is used by organisations (European Commission, 2019).

Access control refers to the mechanisms and policies that regulate who can access and interact with data, systems, and resources. In the context of AI, access control is crucial for protecting sensitive data, ensuring the integrity of AI models, and preventing unauthorized use. It involves defining roles, permissions, and restrictions to limit access based on the principle of least privilege. If not mitigated, failed access control can lead to several adverse situations (Yampolski, 2020):

- Data leaks or breaches
- Biased data access leading to AI models inadvertently inheriting biases present in the data
- Adversarial attacks to manipulate AI models by accessing training data

- Tampering with AI models during the training process
- Model poisoning, where malicious actors access and tamper with AI models during the training process

The following table outlines a risk assessment in the context of data subject rights and access control in AI systems.

Table 25. Risks and mitigation measures in relation to data subject rights and access control in AI systems.

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
R4.1	**Right of access** Failure to ensure safe storage, processing and transfer of requested information responding to data subject requests who wish to receive additional information on the processing of their data, such as details on the information that is kept on them or what it might be used for.	Protocol needs to be established on technology level to always ensure safety of all data handling. Enforce Multi-Factor Authentication (MFA) to enhance authentication security, ensuring that only authorized users can access sensitive data.	Protocol needs to be established on human level to always ensure safety of all data handling. Role-Based Access Control (RBAC) should be implemented to define roles and permissions for users based on their responsibilities, restricting access to data and systems accordingly.	(Hupont et al, 2023) (Yampolski, 2020) (Mueck et al, 2023)
R4.2	**Right to rectification** Failure to update, correct or delete inaccurate or outdated information according to request.	Request for data deletion should be subjected to scrutiny and after verification, follow a strict process with built-in safeguards to prevent such failure (e.g. by issuing alerts).	Request for data deletion should be subjected to scrutiny and after verification, follow a strict process with built-in safeguards to prevent such failure (e.g. by issuing alerts).	(Burgess & Kloza, 2021) (Mueck et al, 2023) (Truby et al, 2022)
R4.3	**Right to erasure** Failure of technical procedures that control the erasure of data of an identifiable individual, in accordance with the internal organisational policies and regulations.	Technology should, by design, ensure that processes are carried out as intended.	Comprehensive inventory of all personal data within systems and processes should be maintained. This inventory helps identify where personal data is stored, ensuring that erasure requests are efficiently addressed.	(Burgess & Kloza, 2021) (Hupont, 2023) (Yampolski, 2020)

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
R4.4	**Proportionality tests** Failure to comply with protocols assuring the proportionality of use as well as safeguards for the protection of freedoms	Algorithms should not infringe or showcase bias according to articles outlined in the EU Charter of fundamental rights. Users are to verify that adequate measures have been taken on the provider end to ensure the system is compliant with the applicable law.	Protocols should be established for such case that flaws in the design process are detected at deployment stage, including definition of required remedial action as way of contingency planning. The balancing exercise between the security and the subject right should be well documented.	(Glauner, 2022) (Ulnicane, 2022) (Yampolski, 2020)
R4.5	**Accountability** System failing to allow users, data protection officers and supervisory authorities to comply with regulation, or failure to adopt updates in that respect, failure to ensure legal compliance.	An error reporting process and "help" infrastructure should be in place which allows authorised users to document and remedy any system error immediately should it occur.	Comprehensive data governance framework that outlines roles, responsibilities, and processes related to data handling, storage, processing and sharing, should be established. An organisation should appoint a Data Protection Officer (DPO) or a designated person responsible for overseeing data protection and privacy compliance.	(Burgess & Kloza, 2021) (Truby et al, 2022) (European Commission, 2019)
R4.6	**Implementation** Failure to ensure data protection by design and by default, or to address privacy and data protection concerns raised by the technology.	Privacy and data protection principles should be integrated directly into the design and development of systems, applications, and processes. These principles should be derived from the risk assessment and privacy impact assessment procedure.	Users are to verify that adequate measures have been taken on the provider end at an early stage to meet demands on data protection and privacy, that this process has been documented and that results of this process are evidenced and reflected in the system. Before deployment, the system should be inspected by independent experts for	(Yampolski, 2020) (Burgess & Kloza, 2021) (Estella, 2023) (Laux et al, 2023a)

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
			scrutiny. Users should ensure that social acceptance, privacy and data protection concerns raised by the technology have been assessed and deemed acceptable in view of law and society's perception.	
R4.7	**Safeguards** Failure to establish safeguards and adequate architecture towards the protection of personal data	Ensure that the system is furnished with adequate safeguards and that architecture has built-in functions to protect personal/private data	Established safeguards should be reviewed and checked by the nominated data security governing body on a regular basis and protocol should be in place for the case that failure of safeguards is being flagged.	(Hupont et al, 2023) (Ulnicane, 2022) (European Commission, 2020c)
R4.8	**Record of processing operations** Failure to store or failure to safely store log all processing of information, interaction of the system.	The logs created from the system should be protected against external changes.	Clear logging policies and procedures should be established to outline what events should be logged, how they should be logged, and the level of detail required.	(Burgess & Kloza, 2021) (Truby et al, 2022)
R4.9	**Access control** Flaw in access control (e.g., on levels of customizable profiles or action traceability mechanisms which are meant to ensure the lawful nature of the processing) and/or failure to prevent unauthorised system access.	Algorithms should inherently be made accessible following the use of authentication system. The algorithms should include authentication and authorisation mechanism inbuilt. Users are to verify that adequate measures have been taken on the provider end to ensure the AI technology is compliant with the applicable law and the	Strict user and role management practices should be implemented, allowing access only to authorised users based on their roles and responsibilities.	(Estella, 2023) (European Commission, 2019) (Hupont et al, 2023) (Burgess & Kloza, 2021)

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
		effective system to keep log records is implemented.		
R4.10	**Control granularity** Error in fine grain access control for different profiles, users, cases which is meant to provide fine granularity at groups, users and cases for access rights permits to information and business processes launch.	Algorithms should support business continuity upon appropriate authentication and authorisation is obtained. The use of industry best practices for adopting single sign on process is encouraged in the design and implementation of algorithms and software components.	Attribute-Based Access Control (ABAC) should be utilized to control access based on various attributes, such as user attributes, data attributes, and environmental conditions. Adhere to the principle of least privilege, granting users only the minimum access necessary to perform their tasks.	(Hupont et al, 2023) (Ulnicane, 2022) (Glauner, 2022

3.5 R5 – Automated Decision-making

Evaluating and addressing the potential risks in policing and law enforcement associated with decisions made by AI systems without human intervention is a critical area in risk assessment and mitigation. This includes assessing the fairness, accuracy, transparency, and potential impact on affected persons. The goal is to mitigate risks and ensure responsible and ethical decision-making processes (Laux et al, Wachter, 2023a).

The process of translating complex data inputs into automated decisions can lead to several harmful outcomes (Hadzovic et al, 2023), such as:

- Discriminatory decisions caused by biases present in training data
- Incorrect decisions caused by subtle manipulations of input data through adversarial attacks
- Mistrust due to the lack of transparency in AI decision-making processes
- Inability to handle uncertain or novel scenarios

AI risk assessment in the context of decision-making involves a systematic evaluation of these potential risks during making automated choices. This assessment entails identifying scenarios where AI-driven decisions might lead to undesirable outcomes, whether due to biases in training data, vulnerabilities to adversarial attacks or the inability to handle novel situations. The following table outlines identified risks and mitigation measures in relation to automated decision-making in policing by AI systems.

Table 26. Risks and mitigation measures in relation to automated decision-making by AI systems.

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
R5.1	**Human error** Human error on automated decision-making level: automated decision-making including profiling, typically requires the approval (and capability of intervention) of a human operator before the results are further progressed into the given system.	Incorporate redundancy and cross-checking mechanisms, where multiple individuals review and verify decisions or critical information before implementation.	Review procedures should be put in place to flag and therefore minimise consequences of such error.	(Burgess & Kloza, 2021) (Truby, 2022) (Hadzovic, 2023
R5.2	**Special categories of data** Failure of blacklist rules which are meant to exclude profiling solely on the basis of sensitive attributes and to ensure that the reasoning process relies on objective and reasonable grounds to avoid discriminatory profiling.	The automated decision-making components should be built on the principle of lawfulness by design. To avoid possibly discriminatory results in the reasoning process, all sensitive attributes like ethnic group or sexual orientation should be excluded from the ontology.	Establish a review process involving ethics committees or designated individuals to assess the ethical implications of using sensitive data for decision-making.	(Hadzovic et al, 2023) (Ulnicane, 2022) (Laux et al, 2023a)
R5.3	**Auditing** Failure of logging mechanisms which are designed to allow for the auditing of system use and/or error in the processes supporting the profiling and automated decision-making practices.	Algorithms implementing the decision-making practices should be capable of being audited for the conclusions drawn. The algorithms should also justify the outcome through statistical quantity.	Conduct regular audits of the decision-making processes involving sensitive data to ensure compliance with policies, regulations, and ethical standards.	(Burgess & Kloza, 2021) (Ulnicane, 2022) (European Commission (2020c) (Laux et al, 2023a)

#	*Risk description*	*Risk mitigation measure suggested - Technical*	*Risk mitigation measure suggested - Operational*	*Sources*
R5.4	**Transparency in decision-making** Lack of transparency of automated process and how outcomes are derived.	Comprehensive documentation of the data used, including sources, preprocessing steps, and any biases present should be maintained. Provide users with clear and understandable explanations of how automated decisions are made.	Establish clear governance policies and practices for AI systems, including decision-making processes and accountability mechanisms. Conduct assessments to understand and document the potential impact of the algorithm on different groups.	(Sanz-Urquijo et al, 2022) (Truby, 2022) (Laux, 2023a)
R5.5	**Safeguards** Lack of review mechanisms for automated process to impose corrective measures when needed to adjust how outcomes are derived.	On technology level, procedures need to be embedded in the use process so that accuracy and functionality of the technology is being reviewed on a regular basis.	On human level, procedures need to be embedded in the use process so that accuracy and functionality of the technology is being reviewed on a regular basis.	(Burgess & Kloza, 2021) (Truby et al, 2022) (Ulnicane, 2022)

4. ALIGNER Risk Assessment Instrument (FOI, CBRNE)

The ALIGNER Risk Assessment Instrument (RAI) aims to help LEAs identify risks related to AI technologies, assess the impact of those risks, and implement relevant mitigations to reduce the likelihood for, or the severity of, risk realisation. The instrument consists of seven templates to help LEAs consider a variety of relevant issues when determining the potential risks posed by use of AI, as well as to help LEAs plan ways of responding to these risks.

Section 2 contained a presentation of already existing instruments for AI technology risk impact assessments. Section 3 complemented section 2 by also introducing mitigation measures that may reduce the likelihood for, or severity of, risk realisation. In section 4, we present the ALIGNER RAI which consists of a selection of the abovementioned risks and mitigation measures that we consider particularly relevant to LEAs. First, in section 4.1 follows some recommendations on how to select people to participate in the risk assessment. section 4.2 consists of a description of the methodology of the ALIGNER RAI. Thereafter follows in section 4.3 the seven templates that together form the ALIGNER RAI.

4.1 Responsibility for Conducting the ALIGNER Risk Assessment

LEAs deploying AI systems (‘users’) are responsible to conduct the ALIGNER RAI. Implementation of the ALIGNER RAI requires multidisciplinary skills: people with technical, legal, and ethical expertise including a wide range of personnel in various units and departments from innovation teams to end users of AI systems should be included in the assessment. It is recommended that the procedure is conducted periodically where interdisciplinary competence is very important. This Risk Assessment is not meant to replace the implementation of any other risk assessment. The instrument takes into account that each law enforcement agency will have its own unique situation where some might already use AI in their day-to-day work, while other agencies are working with getting a general understanding of the available technologies. What is similar to all is the recommendation that the instrument should be conducted at the earliest opportunity and used for the first time prior to the deployment of an AI-system. The procedure should at least be updated when a significant change in the system arises and/or when a new technique is to be implemented.

4.2 Methodology for the ALIGNER Risk Assessment Instrument

The ALIGNER RAI consists of seven templates for risk identification, assessment, and mitigation. Before presenting all seven templates (section 4.3), there follows a description of the method for developing the templates, and an instruction on how to understand and use the templates.

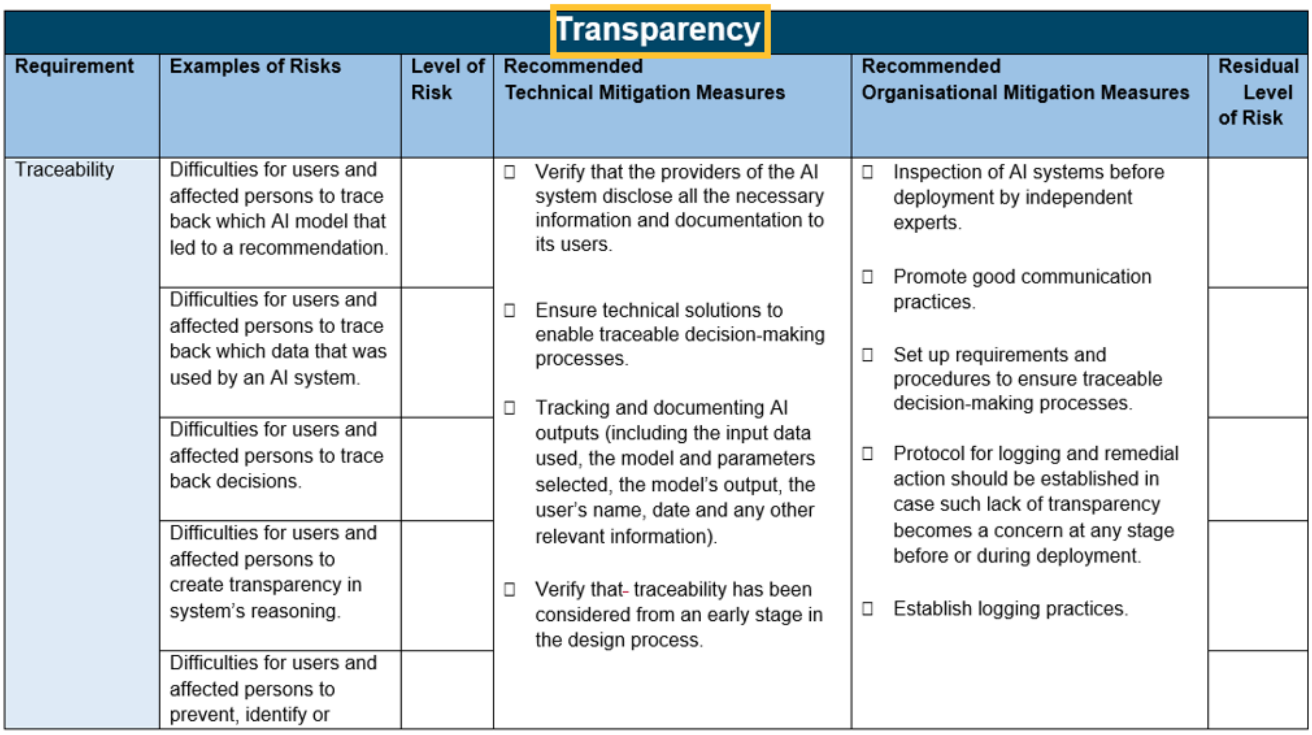
4.2.1 Structure of Risks and Mitigation Measures in the Templates

It is clear from section 2 and 3 above that risks associated with AI technologies can be categorised in many different ways. In the ALIGNER RAI Templates, risks and their related mitigation measures are categorised based on the seven requirements for trustworthy AI, developed by AI HLEG (key requirements) and their associated sub-requirements, as used by the same expert group in the risk assessment instrument ALTAI to specify the key requirements (Section 2.1).

This structure allows for a broad perspective of risks and avoids limiting the risks to those related to data protection or data security for example. This structure is also similar to the one presented in the template for AI system governance described in the deliverable ALIGNER D4.2 (Casaburo and Marsh, 2023). These two templates will therefore naturally complement each other. However, LEAs shall conduct the assessments at different stages in the assessment phase (first the ALIGNER RAI and then continue with the fundamental rights impact assessment (FRIA) as described in D4.2) and with different aims. While the one presented in ALIGNER D4.2 is an ethical and legal impact assessment, the one presented in this report is a technical risk assessment of AI technologies. The FRIA described in ALIGNER D4.2 address topics like fundamental rights (including privacy) and this report will therefore not address requirements related to such rights.

As mentioned, the ALIGNER RAI includes seven templates where each template presents risks and mitigation measures related to each of the seven key requirements for trustworthy AI. In the templates, the heading illustrates which key requirement the presented risks and recommended mitigation measures relate to. (Figure 1).

Figure 1 – Column of Key Requirement marked in yellow.

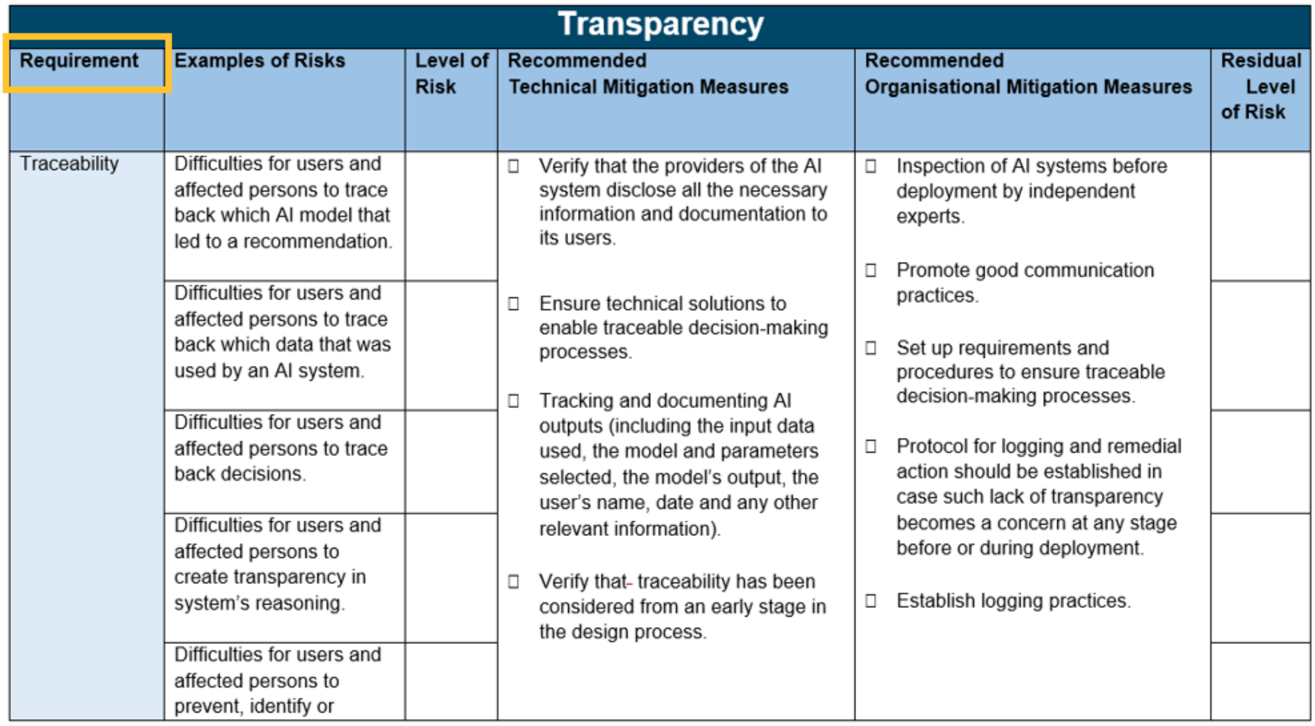


Transparency

Requirement	Examples of Risks	Level of Risk	Recommended Technical Mitigation Measures	Recommended Organisational Mitigation Measures	Residual Level of Risk
Traceability	Difficulties for users and affected persons to trace back which AI model that led to a recommendation.		□ Verify that the providers of the AI system disclose all the necessary information and documentation to its users. □ Ensure technical solutions to enable traceable decision-making processes. □ Tracking and documenting AI outputs (including the input data used, the model and parameters selected, the model's output, the user's name, date and any other relevant information). □ Verify that- traceability has been considered from an early stage in the design process.	□ Inspection of AI systems before deployment by independent experts. □ Promote good communication practices. □ Set up requirements and procedures to ensure traceable decision-making processes. □ Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment. □ Establish logging practices.	
	Difficulties for users and affected persons to trace back which data that was used by an AI system.				
	Difficulties for users and affected persons to trace back decisions.				
	Difficulties for users and affected persons to create transparency in system's reasoning.				
	Difficulties for users and affected persons to prevent, identify or				

Each template also includes a left column with the sub-requirements that relate to each key-requirement (Figure 2).

Figure 2 – Column of Sub-Requirement marked in yellow

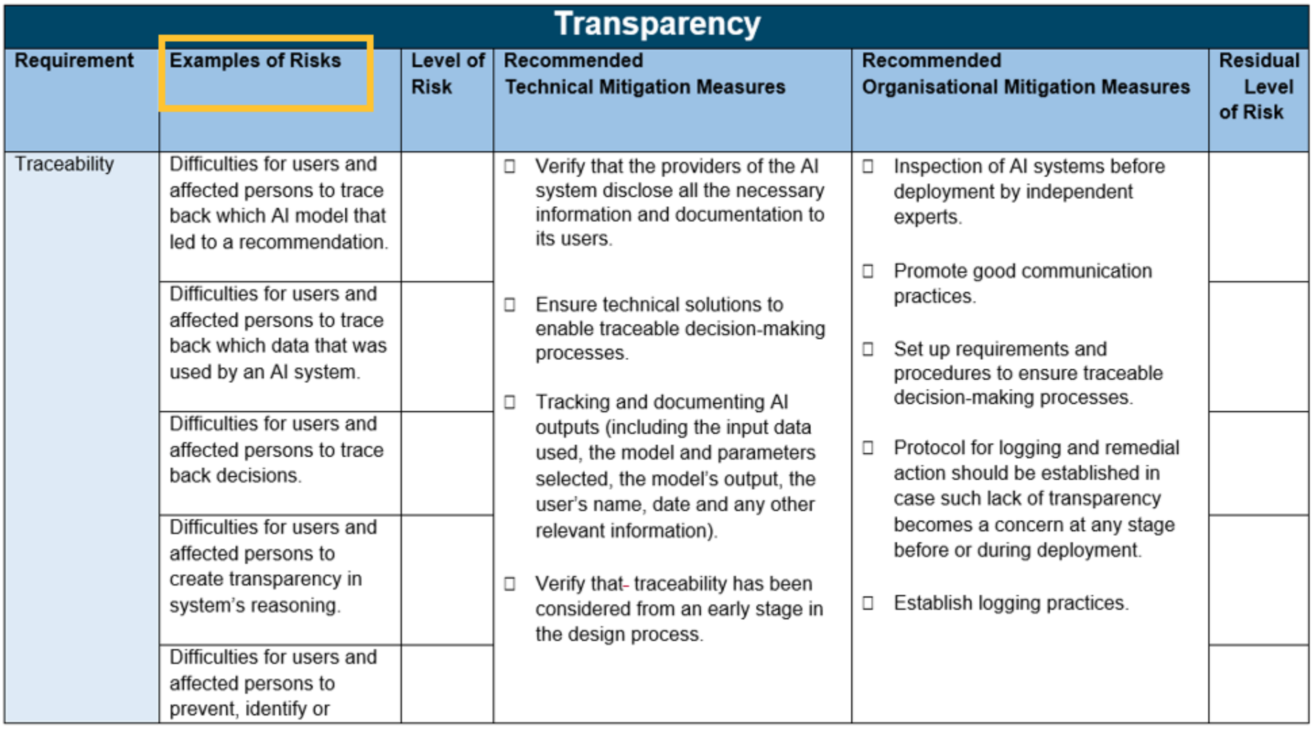


Transparency					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**Residual Level of Risk**
Traceability	Difficulties for users and affected persons to trace back which AI model that led to a recommendation.		□ Verify that the providers of the AI system disclose all the necessary information and documentation to its users. □ Ensure technical solutions to enable traceable decision-making processes. □ Tracking and documenting AI outputs (including the input data used, the model and parameters selected, the model's output, the user's name, date and any other relevant information). □ Verify that traceability has been considered from an early stage in the design process.	□ Inspection of AI systems before deployment by independent experts. □ Promote good communication practices. □ Set up requirements and procedures to ensure traceable decision-making processes. □ Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment. □ Establish logging practices.	
	Difficulties for users and affected persons to trace back which data that was used by an AI system.				
	Difficulties for users and affected persons to trace back decisions.				
	Difficulties for users and affected persons to create transparency in system's reasoning.				
	Difficulties for users and affected persons to prevent, identify or				

4.2.2 The Column "Examples of risks"

To help LEAs to understand the risks of non-compliance with the requirements for trustworthy AI, each template includes examples of risks (Figure 3). These risks are a selection of the risks described further in section 2 and 3, categorised in the structure presented above (section 4.2.1). In the selection of risks, the relevance of the risks for LEAs has been decisive. However, the column of risks is not an exhaustive list of all possible risks. We encourage LEAs to identify and revise risks when relevant for each LEA. Therefore, each template includes empty cells in which LEAs can add any other identified risks not already mentioned in the template.

Figure 3 – Column of Examples of Risks marked in yellow.

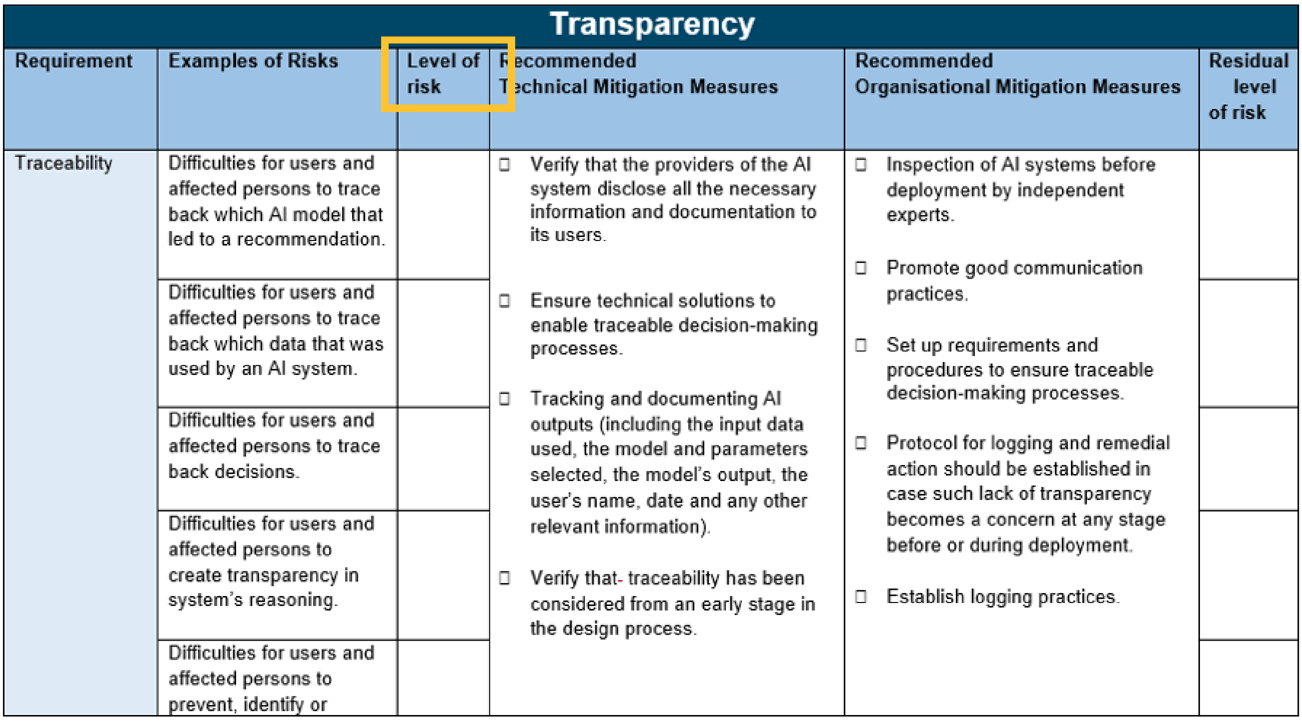


Transparency					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**Residual Level of Risk**
Traceability	Difficulties for users and affected persons to trace back which AI model that led to a recommendation.		□ Verify that the providers of the AI system disclose all the necessary information and documentation to its users. □ Ensure technical solutions to enable traceable decision-making processes. □ Tracking and documenting AI outputs (including the input data used, the model and parameters selected, the model's output, the user's name, date and any other relevant information). □ Verify that- traceability has been considered from an early stage in the design process.	□ Inspection of AI systems before deployment by independent experts. □ Promote good communication practices. □ Set up requirements and procedures to ensure traceable decision-making processes. □ Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment. □ Establish logging practices.	
	Difficulties for users and affected persons to trace back which data that was used by an AI system.				
	Difficulties for users and affected persons to trace back decisions.				
	Difficulties for users and affected persons to create transparency in system's reasoning.				
	Difficulties for users and affected persons to prevent, identify or				

4.2.3 The Column ”Level of Risk”

LEAs should assess each risk to identify the level of risk. The level of risk is defined based on multiplying (a) the likelihood for realisation of risks by (b) the impact of the realisation of risks. The procedure of estimating likelihood and impact of risks is further described below (section 4.2.3.1). The result of this procedure, the risk level, is put into the template in the column "Level of Risks" (Figure 4).

Figure 4 – Column of Level of Risk marked in yellow.



Transparency					
Requirement	Examples of Risks	Level of risk	Recommended Technical Mitigation Measures	Recommended Organisational Mitigation Measures	Residual level of risk
Traceability	Difficulties for users and affected persons to trace back which AI model that led to a recommendation.		□ Verify that the providers of the AI system disclose all the necessary information and documentation to its users. □ Ensure technical solutions to enable traceable decision-making processes. □ Tracking and documenting AI outputs (including the input data used, the model and parameters selected, the model's output, the user's name, date and any other relevant information). □ Verify that- traceability has been considered from an early stage in the design process.	□ Inspection of AI systems before deployment by independent experts. □ Promote good communication practices. □ Set up requirements and procedures to ensure traceable decision-making processes. □ Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment. □ Establish logging practices.	
	Difficulties for users and affected persons to trace back which data that was used by an AI system.				
	Difficulties for users and affected persons to trace back decisions.				
	Difficulties for users and affected persons to create transparency in system's reasoning.				
	Difficulties for users and affected persons to prevent, identify or				

4.2.3.1 Methodology for the ALIGNER Risk Assessment Procedure

The Artificial Intelligence Toolkit (AIT) (UNICRI and INTERPOL, 2023d) inspired this section that aims to help law enforcement agencies to evaluate and prioritize the risks that an AI system may pose. The purpose is to calculate which risks must be managed and in what order, by giving them a risk level score.

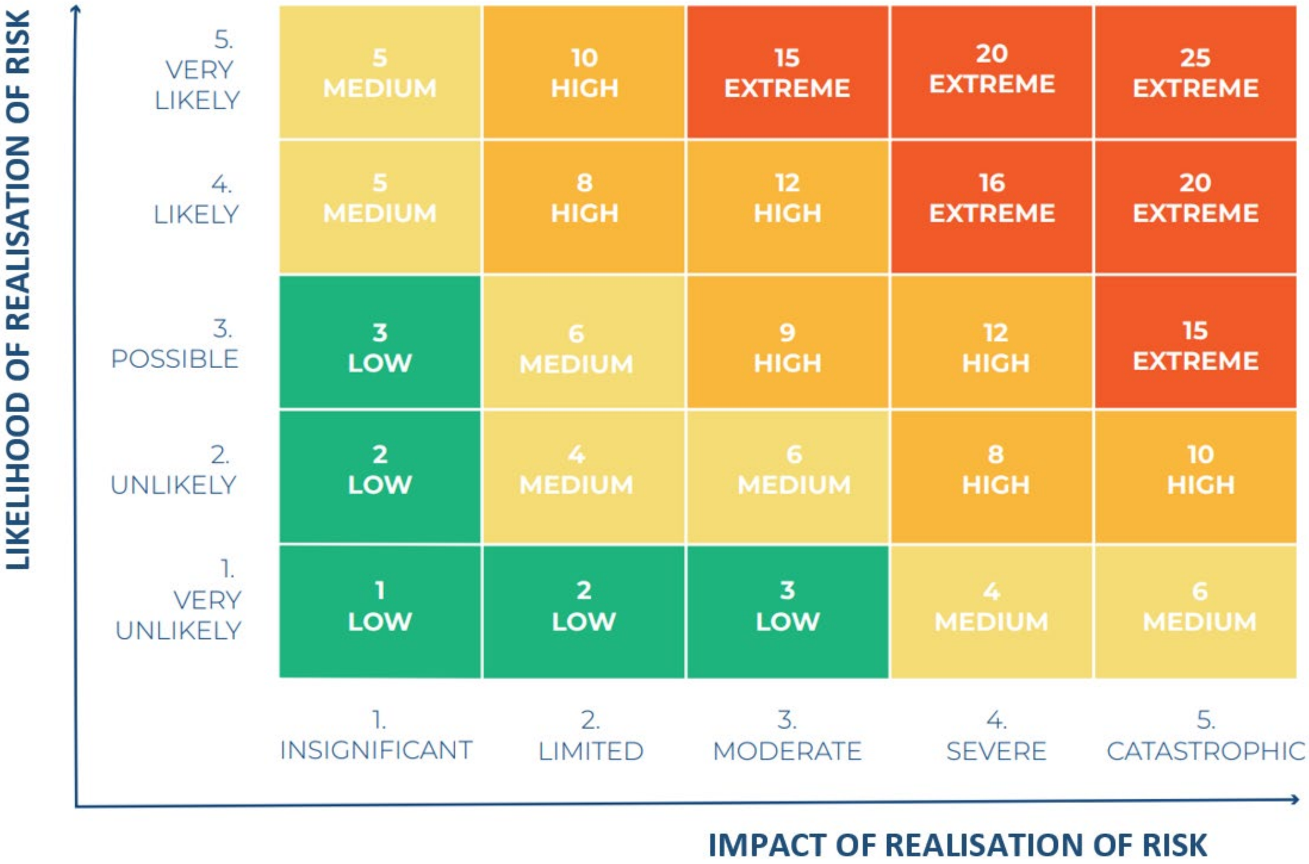
In this report, the assessing and interpreting steps are most relevant for further implementations. The assessing steps aim to evaluate each risk based on two main categories: impact and likelihood. A score of 1 to 5 must be allocated to each of the two main categories. When the evaluation of impact and likelihood for the specific risk have been finalized and scored, the respondent will be able to calculate the overall risk level by multiplying the impact score by the likelihood score for each risk (Level of risk= Impact x Likelihood). The level of risk offers a quantitative measure of the risk, that can be helpful for decision-makers when determining whether to change, approve or withdraw an AI system. It can also be used as a tool to inform and educate users of the adverse impact of an AI system. The score is based on the user's own calculations and opinions of the risks related to a specific AI system.

Figure 5 – Calculate level of risk

Risk	Likelihood (A)	Impact (B)	Level of risk (A x B)
Lack of awareness and insight of the AI system, its use and consequences.	3	3	9

After having determined the level of risk, it is time for the next step, interpreting it. This step aims to generate understanding of the results after determination of the risk score and risk level, with help from a risk matrix and a table of general interpretation. The risk matrix helps the user to visualize and estimate likelihood of realization of risk and impact of realization of risk. The table of general interpretation will help the user to correspond to each risk level by providing a general understanding of the scored level of risk. However, it is important to note that risk matrices can deviate between users from different locales (e.g., while Germany employs a 5x5 risk matrix in civil protection, Spain employs a 6x6 matrix); the interpretations for the risk levels provided below should therefore only be understood as examples.

Figure 6 – The risk matrix (UNICRI and INTERPOL, 2023d)



LIKELIHOOD OF REALISATION OF RISK

	1. INSIGNIFICANT	2. LIMITED	3. MODERATE	4. SEVERE	5. CATASTROPHIC
5. VERY LIKELY	5 MEDIUM	10 HIGH	15 EXTREME	20 EXTREME	25 EXTREME
4. LIKELY	5 MEDIUM	8 HIGH	12 HIGH	16 EXTREME	20 EXTREME
3. POSSIBLE	3 LOW	6 MEDIUM	9 HIGH	12 HIGH	15 EXTREME
2. UNLIKELY	2 LOW	4 MEDIUM	6 MEDIUM	8 HIGH	10 HIGH
1. VERY UNLIKELY	1 LOW	2 LOW	3 LOW	4 MEDIUM	6 MEDIUM

IMPACT OF REALISATION OF RISK

Figure 7 – The table of general interpretation (UNICRI and INTERPOL, 2023d)

	LEVEL OF RISK (A x B)	INTERPRETATION
LOW RISK	1 to 3	The likelihood of the event occurring is very low and/or, if it does occur, the impact will be minimal. Despite being classified as low, these risks should still be managed and prevented or mitigated where possible.

MEDIUM RISK	4 to 6	The likelihood of the event occurring is low, or, if there is a higher likelihood, the impact if it does occur will not be severe and not lead to significant harm. These risks require management that is more active and planning to mitigate.
HIGH RISK	7 to 12	The event is likely to occur, or the impact if it does occur will be severe. These risks require immediate attention and robust mitigation strategies.
EXTREME RISK	13 to 25	The risk is almost certain to occur, or the impact if it does occur will be extremely severe or catastrophic. These risks require urgent, comprehensive action, including redesigning or discontinuing the AI system.

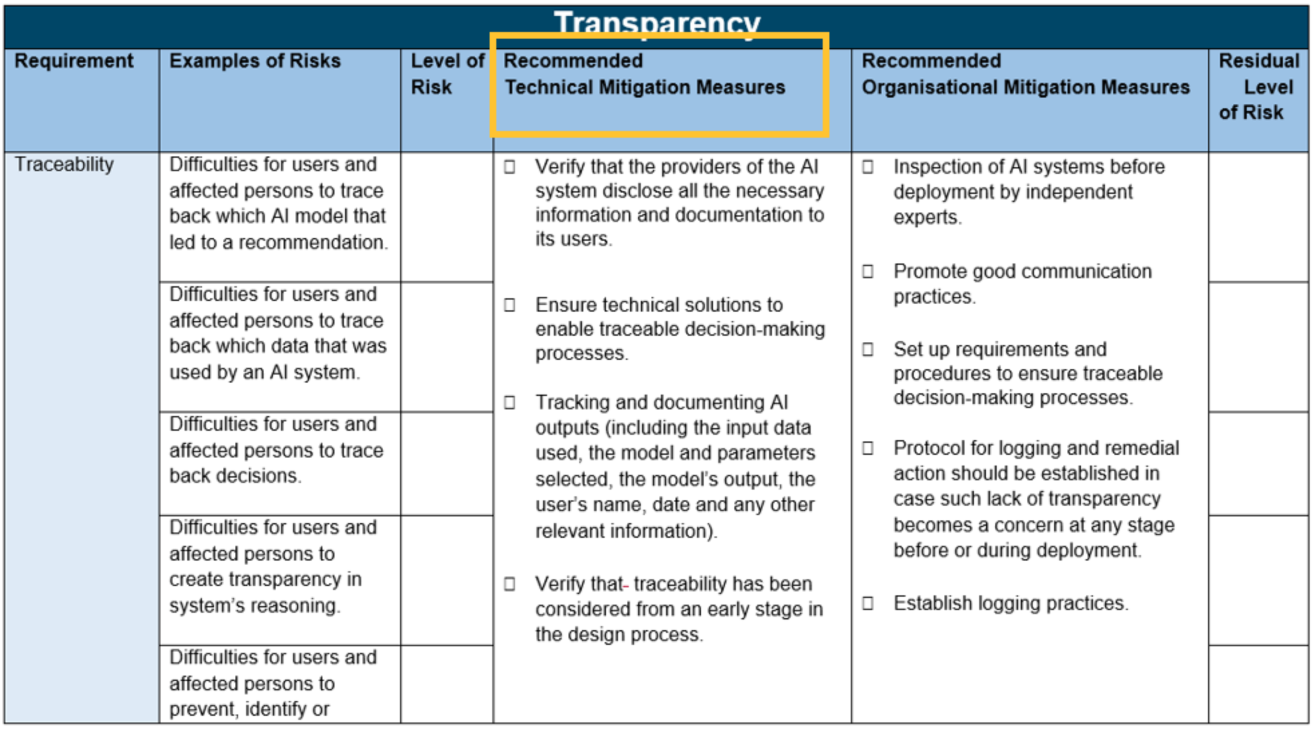
4.2.4 The Column "Recommended Technical Mitigation Measures"

Measures to mitigate risks can take various forms. The templates in the ALIGNER RIA present mitigations measures of two forms: technical and organisational. The technical mitigation measures relate to the AI technology as such. For example, these mitigation measures can correspond to the design and development of AI models as well as test and verification of such models (Figure 8). As shown, we do not recommend mitigation measures for each risk, but for those risks related to the same requirement, as many of the mitigation measures are relevant for more than one of the risks mentioned.

As with risks, this is only a selection of recommended mitigation measures. In addition to these, we encourage LEAs to complement the list with any other relevant mitigation measures than the ones already included in the template. Just as there may be other relevant mitigation measures, not all recommended mitigation measures may be applicable in each case or related to each AI technology. LEAs are encouraged to use the tool in a flexible way.

Some of the mitigation measures we recommend are required by law, while others are not. Like any other activity that LEAs engage in, their engagement with AI must be lawful. LEAs must therefore follow applicable laws at all times. These laws may vary across regions. Any other legally required mitigations than those mentioned here must also be implemented.

Figure 8 – Column of Recommended Technical Mitigation Measures marked in yellow.



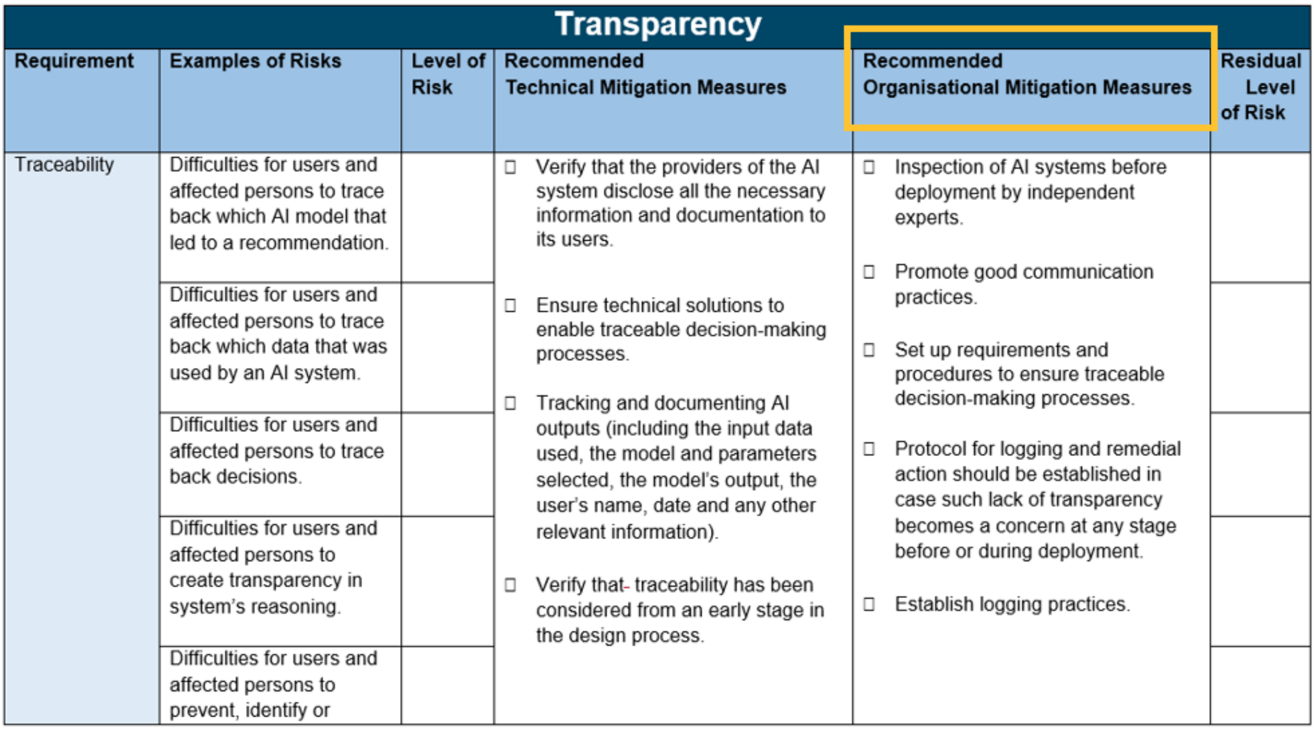
Transparency					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**Residual Level of Risk**
Traceability	Difficulties for users and affected persons to trace back which AI model that led to a recommendation.		□ Verify that the providers of the AI system disclose all the necessary information and documentation to its users. □ Ensure technical solutions to enable traceable decision-making processes. □ Tracking and documenting AI outputs (including the input data used, the model and parameters selected, the model's output, the user's name, date and any other relevant information). □ Verify that- traceability has been considered from an early stage in the design process.	□ Inspection of AI systems before deployment by independent experts. □ Promote good communication practices. □ Set up requirements and procedures to ensure traceable decision-making processes. □ Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment. □ Establish logging practices.	
	Difficulties for users and affected persons to trace back which data that was used by an AI system.				
	Difficulties for users and affected persons to trace back decisions.				
	Difficulties for users and affected persons to create transparency in system's reasoning.				
	Difficulties for users and affected persons to prevent, identify or				

4.2.5 The Column "Recommended Organisational Mitigation Measures"

Even if this deliverable is about technical risks and mitigation, technical and organisational mitigations measures may correlate and be dependent upon each other. Thus, the templates also include organizational measures, namely measures that the organization plans to implement or use the AI technology. For example, these mitigations can correspond to procedures or structures in an organisation (Figure 9). As different LEAs are organised in different ways, the organisational mitigations required in each case may vary. LEAs should take a flexible approach to the recommended organisational mitigations and implement those relevant.

Some of the recommended mitigation measures are required by law, while others are not. Like any other activity that LEAs engage in, their engagement with AI must be lawful. LEAs must therefore follow applicable laws at all times. These laws may vary across regions. Any other legally required mitigation measures than those mentioned here must also be implemented.

Figure 9 – Column of Recommended Organisational Mitigation Measures marked in yellow



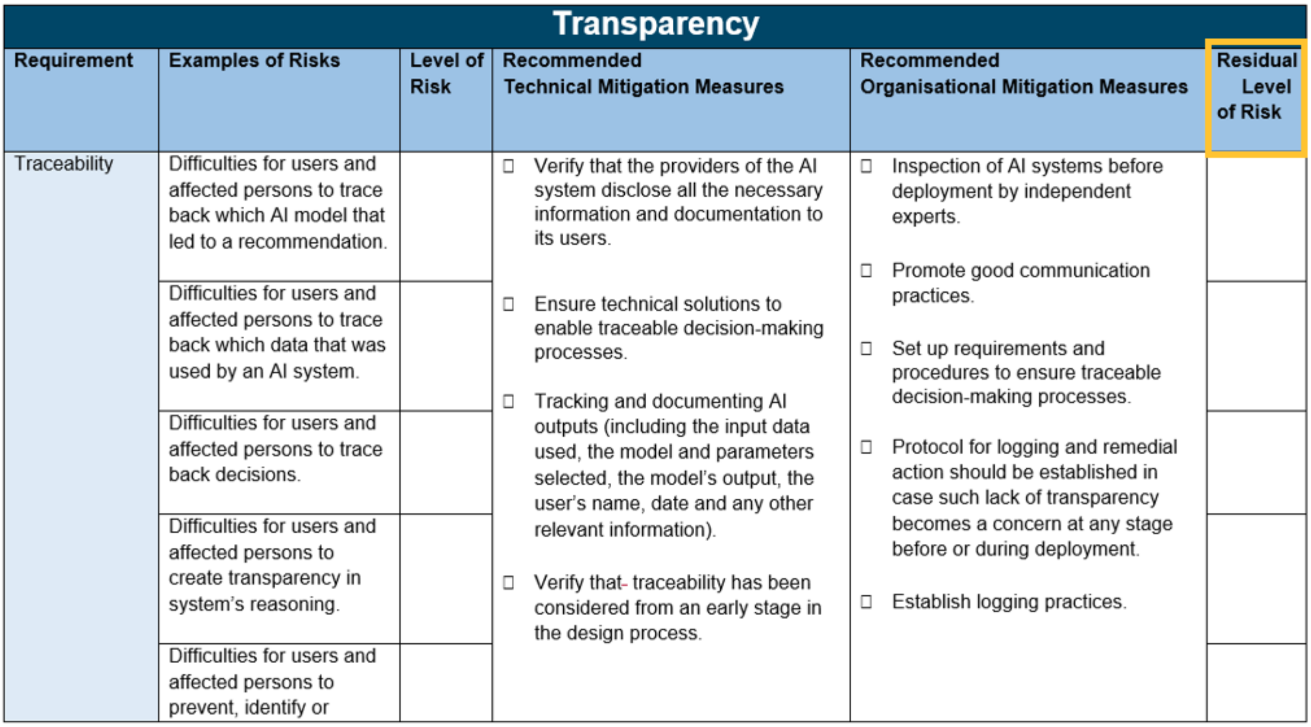
Transparency

Requirement	Examples of Risks	Level of Risk	Recommended Technical Mitigation Measures	Recommended Organisational Mitigation Measures	Residual Level of Risk
Traceability	Difficulties for users and affected persons to trace back which AI model that led to a recommendation.		□ Verify that the providers of the AI system disclose all the necessary information and documentation to its users. □ Ensure technical solutions to enable traceable decision-making processes. □ Tracking and documenting AI outputs (including the input data used, the model and parameters selected, the model's output, the user's name, date and any other relevant information). □ Verify that- traceability has been considered from an early stage in the design process.	□ Inspection of AI systems before deployment by independent experts. □ Promote good communication practices. □ Set up requirements and procedures to ensure traceable decision-making processes. □ Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment. □ Establish logging practices.	
	Difficulties for users and affected persons to trace back which data that was used by an AI system.				
	Difficulties for users and affected persons to trace back decisions.				
	Difficulties for users and affected persons to create transparency in system's reasoning.				
	Difficulties for users and affected persons to prevent, identify or				

4.2.6 The Column "Residual Level of Risk"

The initial risk assessment (section 4.2.3) helps LEAs to assess risks and identify mitigation measures to reduce the risk. After having identified and implemented relevant mitigation measures, LEAs can repeat the risk assessment procedure (section 4.2.3.1). This additional risk assessment will result in a residual level of risk (Figure 10). LEAs can compare the residual level of risk with the initial level of risk to evaluate the effectiveness of implemented mitigation measures. The residual level of risk can also help LEAs to identify the need for further mitigation measures. If LEAs consider the new risk too high, LEAs may also conclude that there is a need to consider alternatives to the assessed AI technology.

Figure 10 – Column Residual Level of Risk



Transparency					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**Residual Level of Risk**
Traceability	Difficulties for users and affected persons to trace back which AI model that led to a recommendation.		□ Verify that the providers of the AI system disclose all the necessary information and documentation to its users. □ Ensure technical solutions to enable traceable decision-making processes. □ Tracking and documenting AI outputs (including the input data used, the model and parameters selected, the model's output, the user's name, date and any other relevant information). □ Verify that traceability has been considered from an early stage in the design process.	□ Inspection of AI systems before deployment by independent experts. □ Promote good communication practices. □ Set up requirements and procedures to ensure traceable decision-making processes. □ Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment. □ Establish logging practices.	
	Difficulties for users and affected persons to trace back which data that was used by an AI system.				
	Difficulties for users and affected persons to trace back decisions.				
	Difficulties for users and affected persons to create transparency in system's reasoning.				
	Difficulties for users and affected persons to prevent, identify or				

4.3 Risk Assessment Templates

In the sections below, we present the seven templates used in the ALIGNER RAI. As this deliverable is focusing on risk assessment of AI technologies from LEA's perspective, the same terminology as was presented in section 1.4 will be used, if not otherwise stated. In column "Recommended Technical Mitigation Measures" and "Recommended Organisational Mitigation Measures", the user (LEA as an agency or its personnel) is responsible for the action.

The sources that has been used to create the templates in section 4 is a combination of the sources used from Section 1-3 (European Commission 2020c, Brey et al., 2020b, Government of Canada, 2023a&b, Secritariat 2023, UNICRI and INTERPOL, 2023a-d, Bostrom 2020, Laux 2023, Lorch 2022, Burgess 2021, Truby 2022, Jacobs 2022, Estella 2023, Hupont 2023, Hadzovic 2023, Sanz-Urquijo 2022 and Glauner 2022).

4.3.1 Human Agency and Oversight

Human Agency and Oversight					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**Residual Level of Risk**
Human agency and autonomy	User or affected persons get confused whether interacting with human or AI system.		□ Verify that the AI systems do not compromise the ability of the users of those systems to act and make decisions independently.	□ Reduce over-reliance among users. □ Training users on how to properly engage.	
	User is over-reliant on AI systems.				

	AI system creating human attachment, stimulating addictive behaviour or manipulating user behaviour.		□ Make user and affected persons aware of when outcomes are a result of an algorithmic decision. □ Inform users and affected persons when they interact with an AI system.	□ Conduct frequent check-ups of AI system that interacts with public to ensure correct functioning.	
	AI system deployed to manipulate and/or control user behaviour.				
	AI system malfunction result in lacking information for user in decision making.				
Human oversight	User lacks training on how to exercise oversight.		□ Verify that mechanisms are established to detect and response to undesirable adverse effects that could affect the user or affected person. □ Procedure for safe abort of operations. □ Verify that the AI systems are built with functionalities that ensure that humans remain in charge (e.g. human in the loop, human on the loop or human in command) during use.	□ Determine how AI system is overseen (in the loop, on the loop, or in command) and by whom. □ Incorporate review procedures.	
	AI systems (especially if making decisions) act without human supervision or intervention.				

				□	
Space for adding own risks and mitigation measures related to human agency and oversight					

4.3.2 Technical Robustness and Safety

Technical Robustness and Safety

Requirement	Examples of Risks	Level of Risk	Recommended Technical Mitigation Measures	Recommended Organisational Mitigation Measures	Residual Level of Risk
Resilience to attack and security	Exposure to cyber-attacks (i.e. data poisoning, model evasion, or model inversion).		□ Certification for cybersecurity. □ Cybersecurity measures. □ Penetration testing. □ Security updates.	□ Measures in place to ensure integrity, robustness, and security against attacks over the AI system lifecycle.	
	Design or technical faults.				
General safety	Damages from technical faults and misuse.		□ Risk identification for technical faults and misuse. □ Definition of safety critical levels. □ Reliability testing. □ Fault tolerance. □ Safety review. □ Use of Transport Layer Security (TLS) protocols to enable secure communication.	□ Process in place to measure and assess risk in each use case (e.g. the ALIGNER FRIA). □ Information to users about risks. □ Safety standards are put in place and updated regularly.	
	Dependency by user on non-robust AI-systems.				
	Insufficient security and system safety standards.				
	Failure to guarantee the ongoing confidentiality, integrity, availability and resilience of processing systems and services.				

Accuracy	Adversarial consequences.		□ Use high quality data. □ Mindful of the origin and composition of the training data. □ Audit of dataset used in training of algorithms. □ Monitoring and verification of accuracy. □ Algorithms validate the input data format against corruption prior to processing the information. □ Testing of system by independent third parties. □ System architecture designed to warn when automatic updates are missed or disabled.	□ Information to user and affected persons about accuracy. □ Training of user. □ Regular "manual" checks to ensure accurate automatic updates. □ Implementation of robust data quality assurance procedures. □ Data quality checks through AI lifecycle.	
	Invalidation of data from operational use				
	Incorrect predictions, recommendations or decisions.				
	Failure of automatic data update mechanism.				
	User dependent on unreliable AI-supported decisions.				
Reliability, fall-back plans and reproducibility	AI system does not behave as expected.		□ Tests to ensure reliability and reproducibility. □ Verification and validation methods. □ Documentation (e.g. logging) to evaluate reliability and reproducibility.	□ Process to monitor if AI system operates as intended.	
	Risks deriving from AI system using online continual learning.				

			□ Fall-back plans. □ Handling of low confidence scores by AI systems. □ Ensure AI systems can perform intended function adequately and cope with changes in its environment. □ The algorithms justify the outcome through statistical quantity.		
Space for adding own risks and mitigation measures related to technical robustness and safety					

4.3.3 Data governance

Data Governance					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**Residual Level of Risk**
Data governance	The processing is not designed to implement by default data protection principles.		□ Data protection principles and standards integrated directly into the design and development of systems, applications and processes. □ Measures to avoid the collection of unnecessary personal data. □ Erasure of no longer necessary personal data. □ Built-in safeguards (e.g. by issuing alerts) to prevent failure to update, correct or delete information. □ Techniques (e.g. differential privacy) to encrypt, anonymise and pseudonymise data.	□ Protocols involving adequate infrastructures and resources should be established by default to ensure compliance with data protection principles, legislation and standards. □ Protocols to ensure compliance with data protection principles, legislation and standards are periodically evaluated and reviewed □ Methods to verify and demonstrate the necessity of the processing for the performance of a law enforcement task. □ Conduct comprehensive inventory of all personal data within systems and processes.	
	Unlawful processing of personal data.				
	Personal data not (further) processed for specified, explicit and legitimate purposes.				
	Unnecessary processing and storage of personal data.				
	The processed personal data are inaccurate or not up to date.				
	Personal data are not anonymised where possible.				
	Unsafe processing, storage and transfer of personal data.				

	Flaw in access control or unauthorised system access.		□ Architecture to detect failure in anonymization and pseudonymisation and ensure remedial action (upon discovery). □ Authentication and authorisation mechanisms embedded in the algorithm. □ Automatic recording of logs. □ Access control mechanisms allowing only authorised users to configure access rights. □ Techniques (e.g. Multi-Factor Authentication (MFA)) to enhance authentication security. □ Combination of purpose-designed system architecture, tailored algorithms, encryption methods and access control measures to prevent unauthorised disclosure of sensitive data. □ Mechanisms (e.g. safeguards) that allow flagging and reporting of data governance issues.	□ Establish regular review for ensuring accuracy and updates of personal data □ Protocols to detect and respond to failure in anonymization and pseudonymization □ Ensure "double up" checks, and alert and flag or forward prompts by combining both human and technology mechanisms. □ Implement Role-Based Access Control (RBAC) to define roles and permissions for user. □ Establish logging policies and procedures to outline what and how events should be logged. □ Protocols to comply with data subject rights. □ Protocols to notify data breaches to supervisory authority and data subjects and to capture necessary detail of the event and its consequences for information or investigation.	
	The data subject cannot exercise their rights.				
	Data breach.				

			□ Protect logs against external changes. □ Ensure "double up" checks, and alert and flag or forward prompts by combining both technology and human level mechanisms. □ Process for regularly testing, assessing and evaluating the effectiveness of the implemented data governance measures. □ AI system developed to allow compliance with relevant data subject rights (e.g. right to access, rectification and erasure).	□ Performance of a data protection impact assessment □ Designation of a data protection officer. □ Training user to comply with data protection principles, legislation and standards. □ Transparent documentation.	
Space for adding own risks and mitigation measures related to data governance					

4.3.4 Transparency

Transparency					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**Residual Level of Risk**
Traceability	Difficulties for users and affected persons to trace back which AI model that led to a recommendation.		□ Verify that the providers of the AI system disclose all the necessary information and documentation to its users. □ Ensure technical solutions to enable traceable decision-making processes. □ Tracking and documenting AI outputs (including the input data used, the model and parameters selected, the model's output, the user's name, date and any other relevant information). □ Verify that traceability has been considered from an early stage in the design process.	□ Inspection of AI systems before deployment by independent experts. □ Promote good communication practices. □ Set up requirements and procedures to ensure traceable decision-making processes. □ Protocol for logging and remedial action should be established in case such lack of transparency becomes a concern at any stage before or during deployment. □ Establish logging practices.	
	Difficulties for users and affected persons to trace back which data that was used by an AI system.				
	Difficulties for users and affected persons to trace back decisions.				
	Difficulties for users and affected persons to create transparency in system's reasoning.				
	Difficulties for users and affected persons to prevent, identify or resolve negative consequences.				
	Lack of awareness and insight for users and				

	affected persons of the AI system, its use and consequences.		□ Ensure that the provider has documented and clearly elucidated all automated processes. □ Implementation of algorithms should adopt transparency and well documented instructions.	□ Ensure that necessary technological and organizational measures are in place.	
Explainability	Difficulties for users and affected persons to understand AI-driven decisions.		□ Information about technical characteristics to enable explainability (e.g., intended purpose, training data set, data sources, potential data set limitation, the level of accuracy etc.) □ The algorithm is designed to enable the user to understand and explain decisions (e.g., InterpretML). □ Algorithms should include enough details on the internal operations to facilitate subsequent data audits. □ The user needs to ensure that the provider has clearly documented all procedures and processes involved in the operation of AI systems.	□ User is to be trained to understand the processes and their impact. □ Continuous survey of if the user understand decision(s) of AI systems. □ Standardize the explanation procedures to ensure consistency and reduce the risk of errors due to miscommunication or misunderstanding.	
Communication	Inadequate communication to user about the AI system's		□ Establish mechanisms to inform about the purpose, criteria and	□ Provide appropriate training and disclaimers to user.	

	limitations and capabilities.		limitations of AI-supported decisions.		
	Difficulties for user and affected persons to understand if interacting with human or AI.				
Space for adding own risks and mitigation measures related to transparency					

4.3.5 Diversity, Non-Discrimination and Fairness

Diversity, Non-Discrimination and Fairness					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigations**	**Recommended Organisational Mitigations**	**Residual Level of Risk**
Avoidance of unfair bias	Inclusion of inadvertent historic bias.		□ Ensure appropriate quality and quantity of training data. □ Consider diversity and representativeness in data. □ Technical tools to understand the data, model and performance. □ Monitoring and reporting of biases during AI lifecycle. □ Verify that the design process was centered around fairness, and results of this processed are evidenced.	□ Inspection of AI systems before deployment by independent experts. □ Structures in place to discover, review and report unfair bias and unequal treatment.	
	Reinforcement of unfair bias that may cause discrimination or inequality.				
	Disadvantage or disproportionately negative impact and harm to certain groups.				
	AI system accessible only by a narrow range of groups and individuals.				

Stakeholder participation	Uninvolved stakeholders			□ Inclusion of stakeholders with regular feedback. □ Long term mechanisms for stakeholder participation in implementation and after deployment.	
Space for adding own risks and mitigation measures related to diversity, non-discrimination and fairness.					

4.3.6 Societal and Environmental Wellbeing

Societal and Environmental Wellbeing

Requirement	Examples of Risks	Level of Risk	Recommended Technical Mitigation Measures	Recommended Organisational Mitigation Measures	Residual Level of Risk
Environmental wellbeing	Negative environmental impacts.		□ Technical measures to reduce carbon emissions during AI lifecycle implemented.	□ Critical examination of resource use and energy consumption during development, deployment and use of AI system.	
	LEA not preserving and improving the welfare of people and the environment in their AI innovation journey as intended or expected.				
	Unsustainable use of resources and high energy consumption.				
Impact on work and skills	AI system alter work sphere in a negative way.		□ Ensure understanding of impacts. □ Counteract de-skilling. □ Technical skill training.	□ Inform impacted workers. □ Skill training.	
	Humans feel their work is not meaningful.				
	De-skilling.				
Impact on society at large or democracy	Negative impact on society at large or democracy.		□ Technical measures to guarantee that AI systems are not harming democratic processes.	□ Examination of direct and indirect societal consequences of AI systems. □ Evaluate if the system could be used for influencing political processes.	

Space for adding own risks and mitigation measures related to Societal and Environmental Wellbeing					

4.3.7 Accountability

Accountability					
Requirement	**Examples of Risks**	**Level of Risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**Residual Level of Risk**
Auditability	Inability to undergo audit.		□ Facilitate auditability, for example by documenting of decisions that influence AI systems' output, and tracking and documenting AI outputs. □ Verify that provider has implemented forensic logs. □ Put in place error reporting process and "help" infrastructure, which allows authorised users to document and remedy any system error. □ Algorithms should include enough details on the internal operations to facilitate subsequent data audits.	□ Conduct regular audits of the decision-making processes.	
	Impossible to supervise the development and use of AI system.				
	Failure to make information on the system programming and functioning available on request.				
	Failure of logging mechanisms which are designed to allow for the auditing of system use.				

Risk management	Irrational tradeoffs occur when implementing risk management.		□ Report, identify and redress risks by design.	□ Facilitate supervision for identifying, assessing, documenting and minimising the potential negative impacts of AI systems. □ Ensure ability to report and respond to the consequences of an AI systems' outcome. □ Ensure protection for entities (e.g. whistle blowers) when reporting concerns. □ Organize risk training.	
Responsibility	Not foreseeing, discover, redressing and reporting risks of AI systems.		□ Mechanisms and processes to enable determination of responsibility. □ The user needs to verify that the provider has taken adequate measures (e.g., documentation) during the design process and accountability.	□ Before deployment, the system should be inspected by independent experts for scrutiny.	
	Responsible persons for decisions made with AI-support not identified.				
	Not redressing the right person.				
	Affected persons cannot challenge AI-supported decisions.				
Space for adding own risks and mitigation measures related to accountability					

5. Risk Assessment of ALIGNER AI Technologies (FOI)

5.1 Risks related to LEA use of techniques in ALIGNER SCENARIO CARDS

The ALIGNER Risk Assessment Instrument (RAI) (section 4) aims to help LEAs identify risks related to AI technologies, assess the impact of those risks and implement relevant mitigation measures to reduce the likelihood for, or the severity of, risk realisation. The instrument consists of seven templates to help LEAs consider a variety of relevant issues when determining the potential risks posed by use of AI, as well as helping LEAs to plan ways of responding to these risks.

In section 5, we present examples of how the ALIGNER RAI can be used together with the ALIGNER Scenario Cards. The examples aim to highlight how the templates may be applied by LEAs. For each scenario card presented, one risk has been identified and described, and one selected template from the ALIGNER RAI has been applied as the figure demonstrates below.

Step one:

Read the scenario card and identify risks in the templates that are relevant in relation to the techniques presented in the card.

Step two:

When the risk has been found in the template, it is time to calculate the level of risk by multiplying impact with likelihood.

Step three:

After defining a risk, it is time to find it in the template and insert the level of risk. By comparing the level of risk with others, it is also time to prioritize which risk that should be assessed.

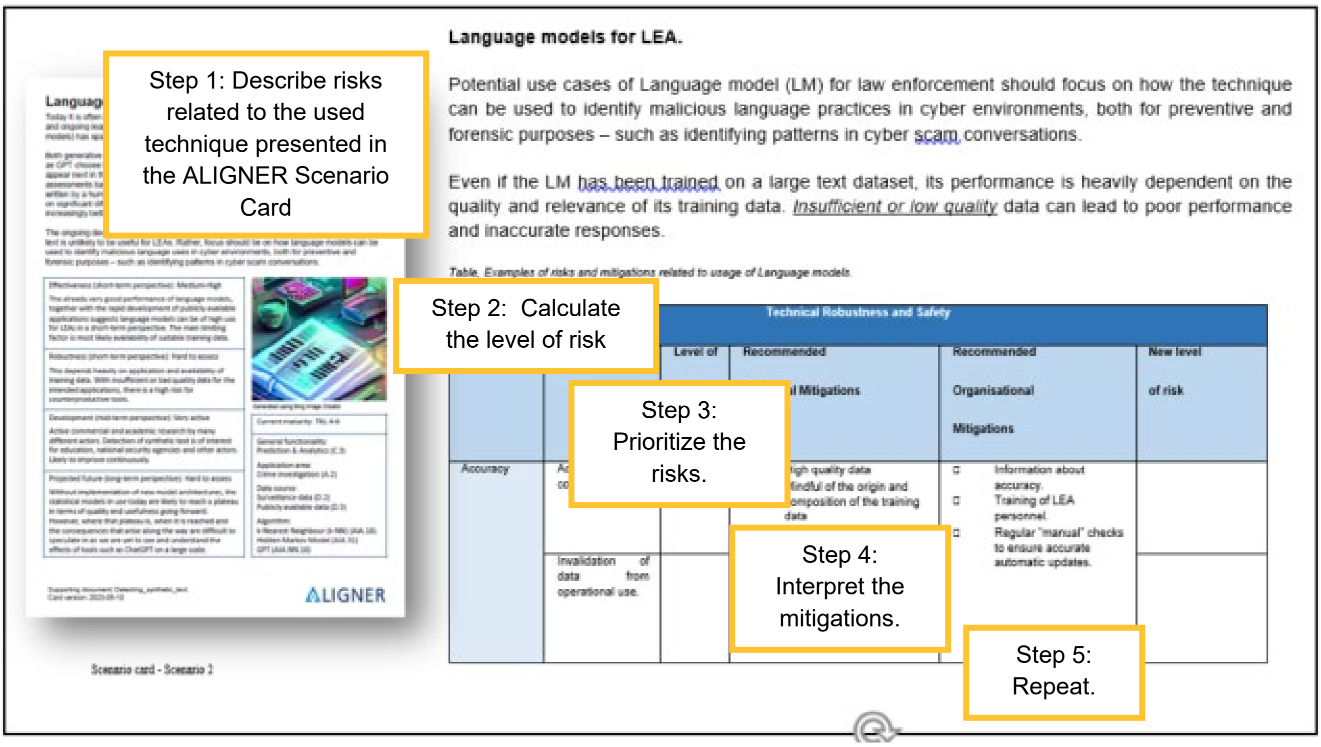
Step four:

After prioritizing, it is time to interpret the mitigation measures for the selected risk.

Step five:

When the mitigation measures have been interpreted, it is time to repeat the procedure to calculate the residual level of risk.

Figure 11: Example of The ALIGNER RAI used together with ALIGNER Scenario Card.



Language

The ongoing de... text is unlikely to be useful for LEAs. Rather, focus should be on how language models can be used to identify malicious language uses in cyber environments, both for preventive and forensic purposes – such as identifying patterns in cyber scam conversations.

Effectiveness (short-term perspective): Medium-High

Robustness (short-term perspective): Hard to assess

Development (mid-term perspective): Very active

Projected future (long-term perspective): Hard to assess

Generated using Bing Image Creator

Current maturity: TRL 4-6

General functionality: Prediction & Analytics (C.3)

Application area: Crime investigation (A.2)

Data source: Surveillance data (D.3) Publicly available data (D.1)

Algorithm: k-Nearest Neighbour (k-NN) (AIA.18) Hidden Markov Model (AIA.31) GPT (AIA.NN.18)

Supporting document: Detecting_synthetic_text
Card version: 2023-08-10

ALIGNER

Scenario card - Scenario 2

Language models for LEA.

Potential use cases of Language model (LM) for law enforcement should focus on how the technique can be used to identify malicious language practices in cyber environments, both for preventive and forensic purposes – such as identifying patterns in cyber scam conversations.

Even if the LM has been trained on a large text dataset, its performance is heavily dependent on the quality and relevance of its training data. *Insufficient or low quality* data can lead to poor performance and inaccurate responses.

Table. Examples of risks and mitigations related to usage of Language models.

Technical Robustness and Safety					
		Level of	Recommended ...l Mitigations	Recommended Organisational Mitigations	New level of risk
Accuracy	Ac... co...		...igh quality data ...indful of the origin and ...omposition of the training data	□ Information about accuracy. □ Training of LEA personnel. □ Regular "manual" checks to ensure accurate automatic updates.	
	Invalidation of data from operational use.				

Step 1: Describe risks related to the used technique presented in the ALIGNER Scenario Card

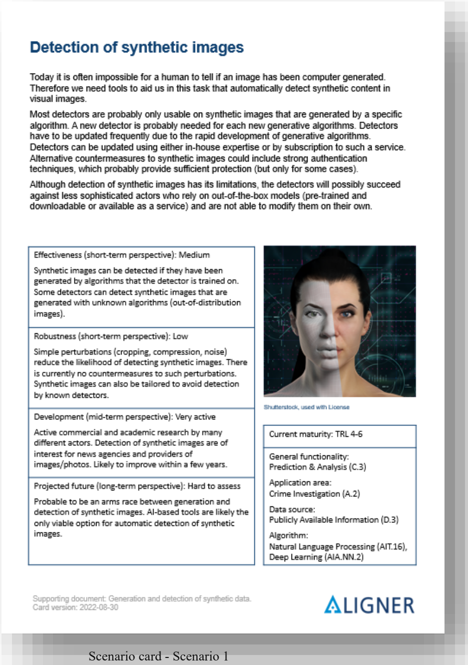
Step 2: Calculate the level of risk

Step 3: Prioritize the risks.

Step 4: Interpret the mitigations.

Step 5: Repeat.

5.1 The ALIGNER RAI used together with ALIGNER Scenario Cards



Detection of synthetic images

Today it is often impossible for a human to tell if an image has been computer generated. Therefore we need tools to aid us in this task that automatically detect synthetic content in visual images.

Most detectors are probably only usable on synthetic images that are generated by a specific algorithm. A new detector is probably needed for each new generative algorithms. Detectors have to be updated frequently due to the rapid development of generative algorithms. Detectors can be updated using either in-house expertise or by subscription to such a service. Alternative countermeasures to synthetic images could include strong authentication techniques, which probably provide sufficient protection (but only for some cases).

Although detection of synthetic images has its limitations, the detectors will possibly succeed against less sophisticated actors who rely on out-of-the-box models (pre-trained and downloadable or available as a service) and are not able to modify them on their own.

Effectiveness (short-term perspective): Medium

Synthetic images can be detected if they have been generated by algorithms that the detector is trained on. Some detectors can detect synthetic images that are generated with unknown algorithms (out-of-distribution images).

Robustness (short-term perspective): Low

Simple perturbations (cropping, compression, noise) reduce the likelihood of detecting synthetic images. There is currently no countermeasures to such perturbations. Synthetic images can also be tailored to avoid detection by known detectors.

Development (mid-term perspective): Very active

Active commercial and academic research by many different actors. Detection of synthetic images are of interest for news agencies and providers of images/photos. Likely to improve within a few years.

Projected future (long-term perspective): Hard to assess

Probable to be an arms race between generation and detection of synthetic images. AI-based tools are likely the only viable option for automatic detection of synthetic images.

Shutterstock, used with License

Current maturity: TRL 4-6

General functionality:
Prediction & Analysis (C.3)

Application area:
Crime Investigation (A.2)

Data source:
Publicly Available Information (D.3)

Algorithm:
Natural Language Processing (AIT.16),
Deep Learning (AIA.NN.2)

Supporting document: Generation and detection of synthetic data.
Card version: 2022-08-30

ALIGNER

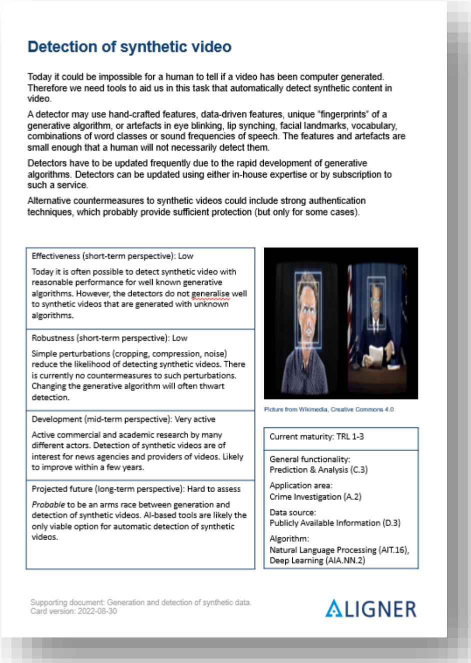
Scenario card - Scenario 1

Detection of synthetic images

Detectors have to be updated frequently due to the rapid development of generative algorithms. Detectors can be updated using either in-house expertise or by subscription to such a service. Alternative countermeasures to synthetic images could include strong authentication techniques, which probably provide sufficient protection (but only for some cases). Detectors can be vulnerable to attacks during which malicious users intentionally input incorrect or misleading information to manipulate the responses of the model.

Table, Examples of risks and mitigations related to usage of Automatic detection of scammer profiles.

Accountability					
Requirement	**Examples of RIsks**	**Level Of risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**New level of risk**
Auditability	Unability to undergo audit.	4	□ Facilitate auditability, for example by documenting of decisions that influence AI systems' output, and tracking and documenting AI outputs. □	□ Define roles and responsibilities.	3
	Impossible to supervise the development and use of AI system.	2		□ Conduct regular audits of the decision-making processes.	1



Detection of synthetic video

Today it could be impossible for a human to tell if a video has been computer generated. Therefore we need tools to aid us in this task that automatically detect synthetic content in video.

A detector may use hand-crafted features, data-driven features, unique "fingerprints" of a generative algorithm, or artefacts in eye blinking, lip synching, facial landmarks, vocabulary, combinations of word classes or sound frequencies of speech. The features and artefacts are small enough that a human will not necessarily detect them.

Detectors have to be updated frequently due to the rapid development of generative algorithms. Detectors can be updated using either in-house expertise or by subscription to such a service.

Alternative countermeasures to synthetic videos could include strong authentication techniques, which probably provide sufficient protection (but only for some cases).

Effectiveness (short-term perspective): Low

Today it is often possible to detect synthetic video with reasonable performance for well known generative algorithms. However, the detectors do not generalise well to synthetic videos that are generated with unknown algorithms.

Robustness (short-term perspective): Low

Simple perturbations (cropping, compression, noise) reduce the likelihood of detecting synthetic videos. There is currently no countermeasures to such perturbations. Changing the generative algorithm will often thwart detection.

Development (mid-term perspective): Very active

Active commercial and academic research by many different actors. Detection of synthetic videos are of interest for news agencies and providers of videos. Likely to improve within a few years.

Projected future (long-term perspective): Hard to assess

Probable to be an arms race between generation and detection of synthetic videos. AI-based tools are likely the only viable option for automatic detection of synthetic videos.

Picture from Wikimedia, Creative Commons 4.0

Current maturity: TRL 1-3

General functionality:
Prediction & Analysis (C.3)

Application area:
Crime Investigation (A.2)

Data source:
Publicly Available Information (D.3)

Algorithm:
Natural Language Processing (AIT.16),
Deep Learning (AIA.NN.2)

Supporting document: Generation and detection of synthetic data.
Card version: 2022-08-30

ALIGNER

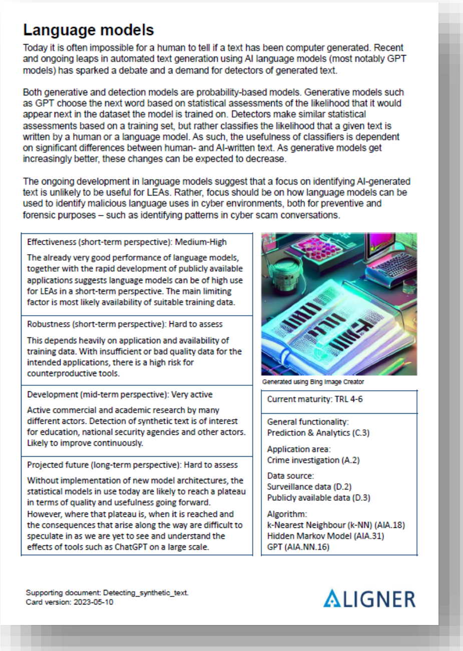
Scenario card - Scenario 1

Detection of synthetic video

Detectors have to be updated frequently due to the rapid development of generative algorithms. Detectors can be updated using either in-house expertise or by subscription to such a service. Alternative countermeasures to synthetic videos could include strong authentication techniques, which probably provide sufficient protection (but only for some cases).

Table, Examples of risks and mitigations related to usage of Automatic detection of scammer profiles.

Technical Robustness and Safety					
Requirement	**Examples of RIsks**	**Level of risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**New level of risk**
Accuracy	Adversarial consequences.	5	□ High quality data □ Mindful of the origin and composition of the training data □ Algorithms should validate the input data format against corruption prior to processing the information.	□ Information about accuracy. □ Training of LEA personnel. □ Data quality checks through AI lifecycle.	4
	Failure of automatic data update mechanism.	3			2



Language models

Today it is often impossible for a human to tell if a text has been computer generated. Recent and ongoing leaps in automated text generation using AI language models (most notably GPT models) has sparked a debate and a demand for detectors of generated text.

Both generative and detection models are probability-based models. Generative models such as GPT choose the next word based on statistical assessments of the likelihood that it would appear next in the dataset the model is trained on. Detectors make similar statistical assessments based on a training set, but rather classifies the likelihood that a given text is written by a human or a language model. As such, the usefulness of classifiers is dependent on significant differences between human- and AI-written text. As generative models get increasingly better, these changes can be expected to decrease.

The ongoing development in language models suggest that a focus on identifying AI-generated text is unlikely to be useful for LEAs. Rather, focus should be on how language models can be used to identify malicious language uses in cyber environments, both for preventive and forensic purposes – such as identifying patterns in cyber scam conversations.

Effectiveness (short-term perspective): Medium-High

The already very good performance of language models, together with the rapid development of publicly available applications suggests language models can be of high use for LEAs in a short-term perspective. The main limiting factor is most likely availability of suitable training data.

Robustness (short-term perspective): Hard to assess

This depends heavily on application and availability of training data. With insufficient or bad quality data for the intended applications, there is a high risk for counterproductive tools.

Development (mid-term perspective): Very active

Active commercial and academic research by many different actors. Detection of synthetic text is of interest for education, national security agencies and other actors. Likely to improve continuously.

Projected future (long-term perspective): Hard to assess

Without implementation of new model architectures, the statistical models in use today are likely to reach a plateau in terms of quality and usefulness going forward. However, where that plateau is, when it is reached and the consequences that arise along the way are difficult to speculate in as we are yet to see and understand the effects of tools such as ChatGPT on a large scale.

Generated using Bing Image Creator

Current maturity: TRL 4-6

General functionality:
Prediction & Analytics (C.3)

Application area:
Crime investigation (A.2)

Data source:
Surveillance data (D.2)
Publicly available data (D.3)

Algorithm:
k-Nearest Neighbour (k-NN) (AIA.18)
Hidden Markov Model (AIA.31)
GPT (AIA.NN.16)

Supporting document: Detecting_synthetic_text.
Card version: 2023-05-10

ALIGNER

Scenario card - Scenario 2

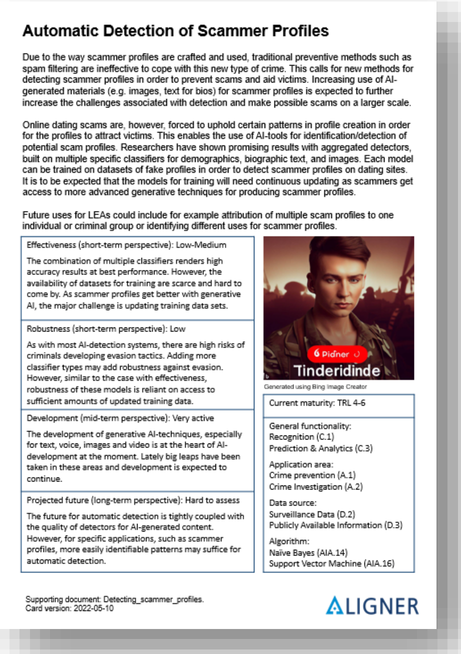
Language models for LEA.

Potential use cases of Language model (LM) for law enforcement should focus on how the technique can be used to identify malicious language practices in cyber environments, both for preventive and forensic purposes – such as identifying patterns in cyber scam conversations.

Even if the LM has been trained on a large text dataset, its performance is heavily dependent on the quality and relevance of its training data. *Insufficient or low quality* data can lead to poor performance and inaccurate responses.

Table, Examples of risks and mitigations related to usage of Language models.

Technical Robustness and Safety					
Requirement	**Examples of RIsks**	**Level of risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**New level of risk**
Accuracy	Adversarial consequences.	6	□ High quality data □ Mindful of the origin and composition of the training data □ Audit of data set used in training of algorithms □ Monitoring and verification of accuracy	□ Information about accuracy. □ Training of LEA personnel. □ Regular "manual" checks to ensure accurate automatic updates.	2
	Invalidation of data from operational use.	7			3



Automatic Detection of Scammer Profiles

Due to the way scammer profiles are crafted and used, traditional preventive methods such as spam filtering are ineffective to cope with this new type of crime. This calls for new methods for detecting scammer profiles in order to prevent scams and aid victims. Increasing use of AI-generated materials (e.g. images, text for bios) for scammer profiles is expected to further increase the challenges associated with detection and make possible scams on a larger scale.

Online dating scams are, however, forced to uphold certain patterns in profile creation in order for the profiles to attract victims. This enables the use of AI-tools for identification/detection of potential scam profiles. Researchers have shown promising results with aggregated detectors, built on multiple specific classifiers for demographics, biographic text, and images. Each model can be trained on datasets of fake profiles in order to detect scammer profiles on dating sites. It is to be expected that the models for training will need continuous updating as scammers get access to more advanced generative techniques for producing scammer profiles.

Future uses for LEAs could include for example attribution of multiple scam profiles to one individual or criminal group or identifying different uses for scammer profiles.

Effectiveness (short-term perspective): Low-Medium

The combination of multiple classifiers renders high accuracy results at best performance. However, the availability of datasets for training are scarce and hard to come by. As scammer profiles get better with generative AI, the major challenge is updating training data sets.

Robustness (short-term perspective): Low

As with most AI-detection systems, there are high risks of criminals developing evasion tactics. Adding more classifier types may add robustness against evasion. However, similar to the case with effectiveness, robustness of these models is reliant on access to sufficient amounts of updated training data.

Development (mid-term perspective): Very active

The development of generative AI-techniques, especially for text, voice, images and video is at the heart of AI-development at the moment. Lately big leaps have been taken in these areas and development is expected to continue.

Projected future (long-term perspective): Hard to assess

The future for automatic detection is tightly coupled with the quality of detectors for AI-generated content. However, for specific applications, such as scammer profiles, more easily identifiable patterns may suffice for automatic detection.

Generated using Bing Image Creator

Current maturity: TRL 4-6

General functionality:
Recognition (C.1)
Prediction & Analytics (C.3)

Application area:
Crime prevention (A.1)
Crime Investigation (A.2)

Data source:
Surveillance Data (D.2)
Publicly Available Information (D.3)

Algorithm:
Naïve Bayes (AIA.14)
Support Vector Machine (AIA.16)

Supporting document: Detecting_scammer_profiles.
Card version: 2022-05-10

ALIGNER

Scenario card - Scenario 2

Automatic Detection of Scammer Profiles

Potential use of AI-tools for identification/detection of potential scam profiles are urgent for LEA. Researchers have shown promising results with aggregated detectors, built on multiple specific classifiers for demographics, biographic text, and images. Each model can be trained on datasets of fake profiles in order to detect scammer profiles on dating sites. It is to be expected that the models for training will need continuous updating.

Like any machine learning model, aggregated detectors, can also be *biased* towards certain groups, topics, or viewpoints depending on the training data it has been exposed to.

Table, Examples of risks and mitigations related to usage of Automatic detection of scammer profiles.

Diversity, non-discrimination and fairness					
Requirement	**Examples of RIsks**	**Level of risk**	**Recommended Technical Mitigation Measures**	**Recommended Organisational Mitigation Measures**	**New level of risk**
Avoidance of unfair bias	Inclusion of inadvertent historic bias.	2	□ Ensure appropriate quality and quantity of training data. □ Consider diversity and representativeness in data. □ Technical tools to understand the data, model and performance.	□ Inspection of AI systems before deployment by independent experts. □ Structures in place to discover, review and report unfair bias and unequal treatment.	1
	Discrimination and inequality.	5			4

References

Asaro, P. M. (2019). *AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care*, in IEEE Technology and Society Magazine, vol. 38, no. 2, pp. 40-53, June 2019, doi: 10.1109/MTS.2019.2915154.

Bostrom, N, Yudkowsky, E. (2020). *The Ethics of Artificial Intelligence and Robotics.* The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, Online resource. https://plato.stanford.edu/entries/ethics-ai/

Brey, P., Lundgren, B., Macnish, K., & Ryan, M. (2020a). *Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach.* De Montfort University. Online resource. https://doi.org/10.21253/DMU.12301322.v1

Brey, P., Lundgren, B.; Macnish, K., & Ryan, M. (2020b). *Guidelines for the Ethical Use of AI and Big Data Systems.* De Montfort University. Online resource. https://doi.org/10.21253/DMU.12301331.v1

Burgess, J. P. & Kloza, D. (2021), *Border Control And New Technologies: Addressing Integrated Impact Assessment*, ASP editions - Academic and Scientific Publishers, doi: 10.46944/9789461171375

Casaburo, D., & Marsh, I. (2023). *ALIGNER D4.2 – Methods and guidelines for ethical & law assessment.* European Commission.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, *9*(13), 1-73.

Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, LED, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016L0680 (accessed on 21 November 2023

Eren, E., Casaburo, D., & Vogiatzoglou, P. (2022). *ALIGNER D4.1 – State-of-the-art reports on ethics & law aspects in Law Enforcement and Artificial Intelligence.* European Commission.

Estella, A., (2023) Trust in Artificial Intelligence: Analysis of the European Commission Proposal for a Regulation of Artificial Intelligence, 30 Ind. J. Global Legal Stud. 39 (2023)

European Commission (2005). *Women and Science: Excellence and Innovation - Gender Equality in Science*. Commission Staff Working Document, SEC(2005) 370, 11 March 2005. Retrieved from https://data.consilium.europa.eu/doc/document/ST-7322-2005-INIT/en/pdf.

European Commission (2019). *Ethics Guidelines for Trustworthy AI: High-Level Expert Group on Artificial Intelligence*. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission (2020a). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. Retrieved from https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

European Commission (2020b). ALTAI portal. Retrieved from https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal.

European Commission (2020c). *White Paper on Artificial Intelligence: A European approach to excellence and trust.* Retrieved from https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Ezeani, G., Koene, A., Kumar, R., Santiago, N., & Wright, D. (2021). *A survey of artificial intelligence risk assessment methodologies – The global state of play and leading practices identified.* Ernst & Young LLP.

Glauner, P. (2022). An Assessment of the AI Regulation Proposed by the European Commission. In: Ehsani, S., Glauner, P., Plugmann, P., Thieringer, F.M. (eds) The Future Circle of Healthcare. Future of Business and Finance. Springer, Cham. https://doi.org/10.1007/978-3-030-99838-7_7

Government of Canada (2023a). *Algorithmic Impact Assessment*. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

Government of Canada (2023b). *Algorithmic Impact Assessment*. https://github.com/canada-ca/aia-eia-js.

Lückerath, D. (2021). *ALIGNER D1.2 – Project Handbook*. ALIGNER – Artificial Intelligence Roadmap for Policing and Law Enforcement. European Commission.

Hadzovic, S., Mrdovic S., Radonjic, M. (2023). *A Path Towards an Internet of Things and Artificial Intelligence Regulatory Framework*, in IEEE Communications Magazine, vol. 61, no. 7, pp. 90-96, July 2023, doi: 10.1109/MCOM.002.2200373

Hupont, I., Micheli, M., Delipetrev, B., Gómez G., Garrido, J. S., (2023). *Documenting High-Risk AI: A European Regulatory Perspective*, in Computer, vol. 56, no. 5, pp. 18-27, May 2023, doi: 10.1109/MC.2023.3235712

Jacobs, M., Simon, J. (2022), *Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed By the European Commission*. DISO 1, 6. https://doi.org/10.1007/s44206-022-00009-z

Krakovna, V., Orseau, L., Ngo, R., Martic, M., Legg, S. (2020). *Avoiding Side Effects By Considering Future Tasks*. Online resource. https://proceedings.neurips.cc/paper/2020/file/dc1913d422398c25c5f0b81cab94cc87-Paper.pdf

Laux, J., Wachter, S., Mittelstadt, B., (2023). *Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act,* Online Resource, http://dx.doi.org/10.2139/ssrn.4365079

Laux, J., Wachter, S. and Mittelstadt, B. (2023), *Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk*. Regulation & Governance. https://doi.org/10.1111/rego.12512

Lorch, B., Scheler, N. Riess, C. (2022). *Compliance Challenges in Forensic Image Analysis Under the Artificial Intelligence Act*, 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 613-617, doi: 10.23919/EUSIPCO55093.2022.9909723

Mueck, M. D., On A. E. B., Du Boispean, S., (2023). *Upcoming European Regulations on Artificial Intelligence and Cybersecurity*. in IEEE Communications Magazine, vol. 61, no. 7, pp. 98-102, doi: 10.1109/MCOM.004.2200612.

OECD (2022). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. OECD.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), GDPR, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679, (accessed on 21 November 2023).

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. AI Now Institute, April 2018. https://ainowinstitute.org/aiareport2018.pdf

Sanz-Urquijo, B., Fosch-Villaronga, E. & M Lopez-Belloso, (2022), *The disconnect between the goals of trustworthy AI for law enforcement and the EU research agenda*. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00235-8

Schiebinger, L., & Klinge, I. (2020). *Gendered innovations 2: How inclusive analysis contributes to research and innovation*. Luxembourg: Publications Office of the European Union.

Schuett, J., (2023). *Risk Management in the Artificial Intelligence Act*, European Journal of Risk Regulation , First View , pp. 1 - 19, DOI: https://doi.org/10.1017/err.2023.1

Secretariat, Treasury Board (2021). Directive on automated decision-making. *Ottawa (ON): Government of Canada (modified 2021-04-01)*.

Secretariat, Treasury Board (2023). *Algorithmic Impact Assessment*. https://open.canada.ca/aia-eia-js/.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), 13-22.

Stahl, B., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B. et al. (2021): D5.8 Artificial Intelligence Impact Assessment - A Systematic Review. De Montfort University. Online resource. https://doi.org/10.21253/DMU.16912387.v1

Truby, J., Brown, R., Ibrahim, I., & Parellada, O. (2022). A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications. European Journal of Risk Regulation, 13(2), 270-294. doi:10.1017/err.2021.52

Ulnicane, I., (2022). *Artificial intelligence in the European Union*, Policy, ethics and regulation, The Routledge handbook of European integrations, DOI 10.4324/9780429262081-19, UNICRI and INTERPOL. (2023a). *Toolkit for Responsible AI Innovation in Law Enforcement: Introduction to Responsible AI Innovation.* Retrieved from: Artificial Intelligence Toolkit (interpol.int)

UNICRI and INTERPOL. (2023b). *Toolkit for Responsible AI Innovation in Law Enforcement: Principles for Responsible AI Innovation.* Retrieved from: Artificial Intelligence Toolkit (interpol.int)

UNICRI and INTERPOL. (2023c). *Toolkit for Responsible AI Innovation in Law Enforcement: Risk Assessment Questionnaire.* Retrieved from: Artificial Intelligence Toolkit (interpol.int)

UNICRI and INTERPOL. (2023d). *Toolkit for Responsible AI Innovation in Law Enforcement: Technical Reference Book.* Retrieved from: Artificial Intelligence Toolkit (interpol.int)

Westman, T., Svenmarck, P., & Chandramouli, K. (2022). *ALIGNER D3.1 – Impact Assessment of AI Technologies for EU LEAs.* European Commission.

Yampolskiy, R., Y. (2020) *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC. https://doi.org/10.1201/9781351251389

Attachments

Explanation of likelihood and impact in the methodology for the ALIGNER Risk Assessment Procedure (4.2.3). The source (UNICRI and INTERPOL. 2023a) has been used to define the different stages of likelihood and impact.

4.4 Likelihood

Likelihood refers to the probability of a certain event or circumstance occurring. In this Risk Assessment, likelihood is defined on a scale of 1 to 5:

1. Very Unlikely: There is a very low chance that the event will occur. It would happen in rare cases or under exceptional circumstances. This scale generally relates to risks that, while possible, are considered negligible.
2. Unlikely: The circumstance is not expected to occur in the normal course of events or frequently. This level generally corresponds to occurrences that, while no longer considered negligible, are uncommon.
3. Possible: There is a fair chance the circumstance or event will occur. It may be triggered by certain conditions or may happen occasionally but the event is not expected to happen consistently or frequently.
4. Likely: The circumstance in the normal course of events. There is a substantial probability that it will occur.
5. Very likely: The circumstance or event is almost certain to occur. It is expected to happen most of the time, barring exceptional circumstances that prevent it (UNICRI and INTERPOL. 2023a).

4.5 Impact

Impact refers to the severity of the potential harm or negative effect that the circumstance or event would have on individuals and communities if it occurred. For the purposes of this Risk Assessment, “individuals and communities” refers to any stakeholder that may be affected by the use of the AI system

1. Insignificant: If the circumstance or event were to occur, it would have minimal or no real impact on individuals or communities. It would not lead to substantial harm or damage but it may cause minor disruption.
2. Limited: If the circumstance or event were to occur, the effects would be relatively contained and manageable but it would cause some damage. It might lead to effort to correct, but it would not cause long-term or widespread harm.
3. Moderate: If the event occur, it would cause a significant level of harm or disruption. This could involve substantial loss of resources or major inconveniences. However, recovery would be relatively straightforward given the right corrective action.
4. Severe & catastrophic: If the circumstance or event were to occur, it would lead to serious harm or disruption. This could involve major losses, significant harm to individuals, severe damage to society or the environment, or considerable legal or ethical implications. Recovery could be difficult, costly, or time-consuming (UNICRI and INTERPOL. 2023a).