ANNEX – POLICY PROTOTYPES

HUMAN OVERSIGHT MEASURES

Knowledge Centre Data & Society

Artificial Intelligence Flanders

Citation: Wannes Ooms, Lotte Cools, Thomas Gils and Frederic Heymans (Knowledge Centre Data & Society), "From Policy To Practice: Prototyping The EU AI Act's Human Oversight Requirements", March 2025

Contact: wannes.ooms@kuleuven.be or thomas.gils@kuleuven.be

www.data-en-maatschappij.ai

Overview

Human Oversight Measures 1

Use case 1: AI education application for student feedback

Human Oversight Measures 1

Contents

1. Introduction

1.1. Purpose of the Document

This document serves as a comprehensive guide for teachers on how to effectively use and oversee the application designed to provide feedback on student writing.

The purpose of this guidance is to:

- Ensure that teachers are equipped with the knowledge and tools to monitor and manage the system's performance.
- Provide clarity on how to interpret and, when necessary, intervene in the feedback generated by the AI system to ensure that it aligns with educational goals.
- Support teachers in preventing over-reliance on the AI system by fostering a critical, hands-on approach to its use.

By following this guidance, teachers will play an essential role in ensuring that the system functions ethically, safely, and effectively in a way that enhances learning outcomes while respecting the integrity of the educational process.

1.2. Overview of the AI application

The AI system aims to provide students with feedback on their assignments, based on assessment criteria determined by teachers. These criteria are central to the application and act as an interface between the neuro-symbolic AI engine that evaluates them and the teaching practices of the educational actors. Students get real-time feedback on their assignments, and teachers get assessment reports that describe how the students are performing across the different criteria in a specific assignment.

The output of the system will be monitored and adjusted to the needs of the teaching staff. If teachers provide feedback about the assessment reports, the system should improve its assessment performance over time. Regarding student privacy, the system will work locally and share only one version of the assessment report with the teacher.

The primary objective of the AI system is not to assign grades to students but to provide them with constructive textual feedback that supports their learning and encourages progress. Assigning grades often shifts students' focus toward the score rather than the learning process, which this system aims to avoid. However, as discussed later in the guidance, some form of measurement is necessary to effectively monitor the AI's performance. To facilitate this, assignments will be internally evaluated with numerical scores (e.g., from 1 to 5) across various dimensions, such as structure, clarity, coherence, etc. These scores will remain undisclosed to both students and teachers and will serve solely for statistical monitoring and system oversight, as explained below.

2. Understanding the AI Educational Application

2.1. The Role of AI in Student Feedback

The goal of feedback is to reduce discrepancies between the current performances of a student and the performances needed to succeed in the task. For feedback to be effective it needs to be concrete and given as soon as possible. It should answer three questions: Where am I going? How am I doing? Where to next?[1]. AI can help students by providing feedback on the task level. The continuous availability of AI makes it possible to provide immediate feedback. AI can also be very specific and detailed, regardless of how many students need to be provided with feedback.

2.2. How the application assesses student writing

The application takes as input the assessment criteria of the teachers and the assignment of the student. Both criteria and user input are translated into logical rules and evaluated for truth using ProSLM[2]: A Prolog Synergized Language Model for explainable domain-specific knowledge-based question answering. The model incorporates a neural translator and a symbolic component that can achieve both goals and then present a report in natural language. Together, they can convey the logical justification that the symbolic component delivers and fact-check the responses given a source of facts.

2.3. Intended Purpose and outcome of the application

The delivery of immediate and meaningful feedback to students to speed up their learning process, respecting their privacy and providing teachers with useful information to improve their practices. Student feedback is delivered locally to the student, and assessment reports are shared with teachers once the student has finished their work. Other actors within the education context may retrieve these reports to inform other educational tasks.

[1] Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

[2] Vakharia, P., Kufeldt, A., Meyers, M., Lane, I., & Gilpin, L. H. (2024, September). ProSLM: A Prolog Synergized Language Model for explainable Domain Specific Knowledge Based Question Answering. In *International Conference on Neural-Symbolic Learning and Reasoning* (pp. 291-304). Cham: Springer Nature Switzerland.

2.4. System's capabilities and limitations

2.4.1. Description of the potential risks of using the AI application

When integrating the application into the evaluation process, several risks should be considered to ensure a smooth and effective experience for both teachers and students. Below are some of the key challenges that may arise during its use:

One of the main risks is the potential mismatch between the AI-generated output and the evaluation criteria. The AI might not always align its feedback with the specific criteria you've established, which could lead to inconsistent or inaccurate assessments of student work. Ensuring that the output meets these criteria can be challenging and may require additional monitoring from teachers.

Another risk is the lack of transparency for students. If the evaluation criteria are not clearly communicated, students might struggle to understand how their work is being assessed, which could lead to confusion or dissatisfaction with the feedback they receive. Clear communication is essential to give students a transparent view of the process.

There is also a possibility that the feedback generated by the AI may not make sense or fail to align with the original evaluation criteria. This could result in students receiving inaccurate or irrelevant feedback, which may negatively impact their learning. To mitigate this, it is advisable to allow students to flag feedback they believe is incorrect. However, this introduces the risk of misuse, where students may flag feedback too frequently or unnecessarily, overwhelming the system and adding to the teacher's workload.

In addition, the process of reviewing flagged content poses its own challenges. Teachers will need to verify whether the flagged feedback is justified, which can be time-consuming and increase their responsibilities. If students overuse the flagging system, this can further complicate the process, requiring teachers to spend more time managing feedback than intended.

Finally, there is the risk of bias within the evaluation criteria themselves. If the criteria are biased, this could affect the AI's feedback, leading to unfair assessments. Teachers should regularly review the criteria to ensure they are objective and fair for all students, as biased criteria can have a significant impact on how students perform and are evaluated.

By being aware of these risks, teachers can take steps to mitigate them and ensure that the AI application supports a fair and transparent evaluation process.

2.4.2. Description of potential misuses

While the AI system offers valuable support in the evaluation process, it also presents opportunities for misuse that could undermine its intended purpose. Two significant risks in this regard involve how both management and teachers may utilize the system in ways that go beyond its original goals.

First, there is the risk that management might use the AI system as a tool to control and supervise teachers. Although the system is designed to assist in student evaluation, it could be misappropriated as a way to monitor teachers' performance, teaching methods, or compliance with institutional guidelines. This could lead to unnecessary pressure on teachers, turning the system into a surveillance mechanism rather than a supportive tool. Such misuse might undermine the trust between teachers and management, and potentially shift the focus from improving student outcomes to scrutinizing teacher behavior.

Second, teachers themselves might misuse the AI system by relying on it as a primary grading tool, rather than using it as a support mechanism. While the system can help in providing feedback or assisting with evaluation tasks, its purpose is not to replace the teacher's role in assessment. If teachers rely too heavily on the application for grading, this could reduce the depth and quality of personalized feedback students receive. Over-reliance on automated systems risks overlooking the nuanced understanding and judgment that teachers bring to the evaluation process, thereby compromising the educational value of the assessments.

3. Measures to Ensure Human Oversight

3.1. Monitoring System Operation

3.1.1. Recognizing Anomalies, Dysfunctions, and Unexpected Outputs

To ensure proper operation of the AI system, statistical monitoring tools will be implemented to analyze anonymized student performance trends over time. These tools will be used to detect anomalies, dysfunctions, or unexpected outputs. Any significant deviations in anonymized performance data, inconsistencies in feedback, or large variations in the application of assessment criteria will be flagged for review. Additionally, teachers will periodically be able to review a selection of anonymized student assignments along with the internal scores assigned by the AI system. This allows educators to verify the coherence and accuracy of the AI's scoring and to ensure that the feedback provided aligns with their expectations. If discrepancies are found, further investigation and adjustment of the system will be conducted.

3.1.2. Monitoring the AI's learning process

The AI's learning process, influenced by teacher feedback, will be continuously monitored to ensure alignment with educational goals. As the system adjusts its assessment methods, anonymized student data will be analyzed to ensure that the changes lead to improved accuracy and fairness. Periodic reviews by teachers of how the system assigns scores to anonymized assignments will also help validate the consistency of the AI's scoring model. Key performance indicators (KPIs), such as assessment consistency across assignments and alignment with teacher expectations, will be reviewed. Human supervisors will oversee these processes to ensure that the AI's learning continues to reflect pedagogical standards and that anonymized student progress is evaluated in a fair and meaningful way.

3.2. Understanding the system's output

3.2.1. Interpretation of feedback reports

In a startup phase, it is probably a good idea to exclude teachers who are not very equipped to work with an algorithm (non-tech savvy teachers). This is to prevent the system from being incorrectly used and evaluated because lack of general knowledge of how these kinds of systems work.

In order to correctly interpret the feedback, several measures can be applied although none of them go without adding an extra risk.

- The statistical monitoring described above can be used as a way to interpret the feedback reports. However, this implies that a (basic) scoring mechanism depending on the sample of students and the kind of assignment it can be hard to anonymise this data completely
- A report button can be a useful tool for students to flag feedback that contradicts the evaluation criteria
- The interface itself should come with instructions on how to use it. This should help both students & teachers how they can interpret the feedback reports.

3.2.2. Explanation of criteria used by the system

It will have to be made clear to the students which criteria the model uses to make the evaluation and how it applies these in the feedback. The logic engine should be explainable to the students/teachers so that they understand what actions they can take to improve their work. If the criteria itself and the way they are applied are a black box the model might miss the goal of giving useful feedback.

3.3. Managing Automation Bias

The system promotes the educational agency of its users thanks to the natural language explanation of the reasoning process that it followed to determine if the assignment passes each criterion. Moreover, if there is a disagreement with this output, teachers can deliver feedback to improve the system and students can complain about their grades. These three mechanisms give control back to the educational actors as conflicts are made explicit and dealt with by humans.

3.4. Instructions for the flagging system

Using the system may contain the risk that given feedback does not make sense or does not align with the initial evaluation criteria given. To prevent this from happening and to allow learners to signal any anomalies in the feedback, students will be able to flag incorrect or weird content. This will be made possible by a report button, visible in the tool. The flagging signal will be sent to the teacher so he or she can check if the feedback is indeed incorrect. In this way, a double human check will be built into the system. Once a teacher indicates flagged feedback as incorrect, this feedback will be given back to the system so it can learn from these mistakes.

4. Intervening and adjusting the system

4.1. Disregarding, overriding, or reversing AI feedback

When feedback for a specific task or assignment is repeatedly flagged by multiple students, it may indicate an underlying issue with the AI's assessment that warrants further investigation. In such cases, the flagged feedback will be reviewed by teachers and may potentially be overridden or reversed if found to be consistently inappropriate or unhelpful.

If a single student frequently flags feedback, this will prompt a closer examination to determine if there is a particular issue affecting that student's interaction with the system. Teachers will assess whether the feedback for that student is problematic or if other contextual factors are at play.

4.2. Deciding when not to use the system's feedback

The model will need to be continuously overseen. If the human that is overseeing this noticed that the feedback mechanism is no longer giving feedback according to the criteria this could lead to a decision to stop using the AI system. Similarly when all feedback is flagged by the students or the statistical monitoring indicates an illogical trend this might be an indication that something is off and the model should (temporarily) be disabled.

Other reasons to stop using the system's feedback might be potential misuse of the system. For instance, if the teacher has found a way to use the system to grade the students instead of just having a helping hand to give them formative feedback.

4.3. "Stop Button" and System Halt Procedures

The system will offer a user interface to manage and stop the service that runs the system in the learning management system of the educational institution. This way, it is possible to stop students and teachers from using the system if deemed necessary by the relevant authority.

5. Support and Resources

5.1. Contact and Support Channels

If you have any questions, or concerns, or need assistance while using our AI application, our dedicated support team is here to help:

- **Customer Support**: +32 XXX.XX.XX.XX
- **Email Assistance**: support@company.com
- **Online Portal**: www.company.com/support

Our support team is available from 9:00 AM to 6:00 PM (CMT) Monday through Friday. We strive to respond to all inquiries within 24 hours

.

Human Oversight Measures 2

Use case 2: Cardiovascular Imaging

Human oversight measures 2

Contents

1. Use Case

1.1. Description

Medical imaging, in particular ultrasound imaging of cardiac microvasculature systems, typically renders a lower resolution image whereas the actual raw signal contains much more data. While traditional imaging systems are not capable of rendering higher resolution images, rapid progress is being made in artificial intelligence supported approaches. In these cases, an AI system is capable of generating higher resolution images as well as detect specific anomalies within the data. In addition, the AI system could also provide an automatic diagnosis to support the cardiologist's diagnosis.

1.2. Purpose

The purpose of this application is twofold:

- Provide high resolution images which allow cardiologists during their diagnostic process
- Support cardiologists in decision making through automatic diagnosis

1.3. Risks

When introducing an AI system for medical imaging, automation bias is a large risk. Within the context of this case, cardiologists are very limited in time. When an AI aid is introduced which has the intention to support and facilitate diagnosis, overreliance on the AI's results is likely to occur.

Additionally, training AI systems for medical use is very dependent on the specific application. For example, a system trained on detecting anomalies in one organ cannot be transposed to any other organ. Also, specific parameters regarding patient characteristics can dramatically influence the type or analysis required.

2. Risk mitigation measures

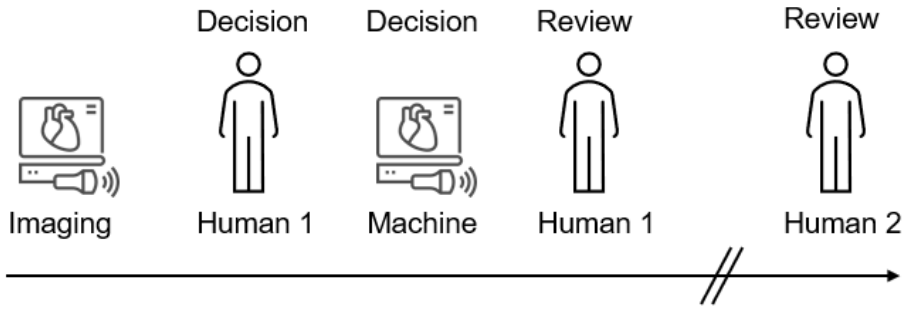
2.1. Human decision making

In order to promote human-centered decision-making and mitigate excessive reliance on automation, the system employs a tiered approach in reaching its conclusions. Following rules are implemented:

- Where the system has arrived at a set of conclusions, such conclusions shall be disclosed only after the human user has submitted their own findings.
- For each conclusion reached by both the system and the human user, the system assigns a probability score and duly communicates these scores to the human user.
- The human user may revise his first decision based on the additional information provided by the system.

- The initial, system-generated, and revised decisions shall be subject to review and approval by a second human before the diagnostic process is finalized.

The rules impose a fixed order of operations illustrated in Figure 1.



A decision shall be deemed well-supported only when the conclusions reached by the human-machine-human sequence are in alignment; specifically, when the determinations of the first human, the machine, and the second human concur. All other decisions shall be deemed inadequately supported and shall be flagged as such by the system.

2.2. Review mechanisms

2.2.1. Technical

To ensure that the AI system functions optimally and that any technical anomalies, inefficiencies, or unexpected results are addressed promptly, the system contains a mechanism for end-users (i.e. cardiologists) to give feedback directly to the provider of the AI system. This feedback process includes the following elements:

- **In-App Reporting System:** Users have access to a simple and intuitive interface for reporting technical issues, anomalies, or unexpected behavior within the system itself. This reporting mechanism allows users to submit logs or specific details related to the issue, as well as attach relevant imaging data, when appropriate.
- **Regular Feedback Channels:** The AI system also includes a functionality that encourages regular, non-urgent feedback from users to inform future updates and refinements. For example, a periodic survey or feedback form could help gather insights about system performance, usability, and potential improvements from experienced medical practitioners.
- **Secure Storage and Extraction of Feedback:** To ensure the security and integrity of the system, it is not necessary to connect the system to the internet. Reported items and user feedback are stored locally and will be reviewed by a technician upon the following routine check of the systems. If necessary, the technician then has the option to securely extract the files for further assessment. The deployer may opt for functionalities with more connectivity if deemed necessary.

The provider commits to timely responses and updates when feedback is received, including providing users with information on corrective measures or system updates made in response to the feedback.

2.2.2. Medical

In addition to the above tiered human decision-making approach, the system as well as its outputs and the diagnoses by the cardiologist should be regularly peer reviewed by other medical experts to ensure the accuracy of the AI system and the correct use of the system by cardiologists. These reviews can also be performed by medical students in the late stages of their education, provided that appropriate nuance is given to their findings.

2.3. System usage

2.3.1. Instruction Manual

A comprehensive technical user manual is crucial to ensure that the AI system is used safely, effectively, and in line with its intended purpose. The system includes a user instruction manual with clear instructions on how the AI system works, how it should be used, and what limitations it has. It includes the following aspects:

- **System Overview and Purpose***:* A clear explanation of the AI system's role in supporting diagnostic decisions regarding cardiac microvasculature using super-resolution imaging with ultrasound. It emphasizes that AI is a support tool and should not be used to replace clinical judgment.
- **Step-by-Step Instructions**: Detailed, user-friendly instructions on how to operate the system, including:
    - How to upload and analyze ultrasound images
    - How the AI system processes data and produces results
    - How the system is best calibrated to provide the most accurate results
    - How to interpret the system's probability scores and findings
    - How to review and validate the AI's conclusions in conjunction with human findings, including an explanation of the aforementioned human decision making process
    - Specific warnings against using the system outside its designed scope, such as on other body parts or on patients unsuited for the technology
- **System Limitations:** The manual includes details on known limitations of the AI system, such as potential edge cases where the system may produce lower confidence results or should not be relied upon
- **Troubleshooting and Reporting:** The manual includes Instructions for resolving common technical issues, as well as steps for reporting bugs, anomalies, or performance problems to the provider. This section should reinforce the feedback mechanisms outlined previously.

The manual will be regularly updated as the system evolves and as new features or improvements are introduced. It is therefore advised to regularly revisit the manual.

2.3.2. Recommended User Profile

By ensuring that the AI system is operated by highly qualified professionals, the risks associated with misuse or over-reliance on the technology are minimized, and the quality of care provided to patients is upheld. Given the sensitivity and complexity of the diagnostic process, the recommended ideal user profile contains the following criteria:

- **Clinical Expertise:** The user should be a cardiologist or, at a minimum, a medical professional with extensive experience in cardiovascular imaging and diagnostics. This expertise ensures that the AI system is used as intended—to support and enhance human expertise, not replace it.
- **Familiarity with Ultrasound Technology:** The user should have advanced knowledge of ultrasound technology and cardiac microvasculature to accurately interpret the resulting images.
- **AI-Specific Training:** Users must have followed specific training on how to operate this AI system, including how to interpret its outputs and understand its limitations. This training should emphasize the risks of automation bias and encourage active engagement with the AI's results rather than passive acceptance.

2.4. Output generation

In order to keep traceability and logging of all automated and AI supported analysis as complete as possible, it is advised to include specific measures regarding the output this system generates.

2.4.2. Watermarking

Any image created must have at least two types of watermarks:

- A visible watermark or label on the exported images, stating that the image being consulted has been improved by an AI system.
- A watermark part of the output's metadata, which contains all information relevant to the processing done by the AI system. This must contain an overview of system state and time of use as well as the specific alterations or manipulations done.

2.4.3. Revertability

All files processed by the system must be able to be reverted to their original state. This implies when an AI system enhanced a specific file, it must be possible to 'roll back' changes. In this way a full history of all actions is kept, very similar to how a versioning system works in software development (e.g. GIT[3]).

[3] https://git-scm.com/

2.4.4. Preconditioning

Specific preconditions for system operation must be taken into account. These preconditions may include specific training data such as racial parameters or relating to biological aspects of patients. Should these preconditions not be met, the system should halt. In this case, the system may be overruled only when a specific logging is made (see watermarking).

Human Oversight Measures 3

Use case 3: Big Data Policing

Human Oversight Measures 3

Contents

1. Use Case

1.1. Description

Big data policing is an innovative strategy that uses historical data to forecast when and where there is a high risk of new crime events (residential burglaries) in order to use police resources more efficiently and proactively, and ultimately reduce crime rates. Big data policing models can consist of variables based on crime data available in police databases (e.g. previous crime events), socio-economic data (e.g. poverty index, residential mobility), opportunity characteristics (e.g. the presence of shops, distance to the nearest highway), data from new technologies (e.g. intelligent cameras) and other known predictors of crime (e.g. police patrol intensity[4]).

2. Human oversight measures

One measure to facilitate human oversight is the production and publication of a (non-technical) **infographic** about the factors, data, metrics and/or features that are used by the big data policing system when providing a specific risk prediction or route recommendation. This infographic could take the form of a dynamic interface/dashboard or of a static publication (e.g. a leaflet). The second human oversight measure that was developed is a **training course** for police officers.

2.2. Infographic

2.2.1. Content

- The Infographic should include information regarding the features that are used by the system to generate risk predictions or route recommendations. An example feature is the density of shops. If this feature would change, then the infographic would also have to explain how such changes influence the prediction. In addition, such features are area-specific. Therefore, the infographic should account for the fact that the spatial combination of features changes and impacts the risk predictions and related route recommendations.

  Other features could be a poverty index, residential mobility, population density, and employment rates, presence of schools, public transport, distance to highways, availability of entry/exit routes, police patrol intensity, previous arrest records, police response times, …

- This infographic could also share layered information about the envisaged impact and related KPIs of the system as well as information about how related concerns are handled.

    - E.g. information regarding how to report issues with regard to the predictions or recommendations, or more generally, the AI system. This

[4] This wording was used in the version of the prototype document provided for feedback. The use case provider has since let us know that this is an inaccuracy. The AI-system in the use case does not use interventions or policy patrol intensity as predictors of crime. However, interventions can occur as a result of the system's predictions.

should build on an organizational overview, including a related hierarchy and the agency by different actors (The police zone and officers versus the local lawmakers versus the AI provider, etc.) It is important to provide instructions on how stakeholders can report errors or inaccurate route suggestions and contribute feedback to improve the system.

    - Prediction Accuracy: How often does the system correctly predict crime risk? This could be represented by percentages or comparisons with actual outcomes.

    - Feedback Effectiveness: Does feedback lead to system improvement?

    - Model Drift Detection: KPIs that monitor how often the system needs recalibration due to changes in underlying data (e.g., new crime patterns, shifts in neighbourhood composition, …).

- The infographic must be updated regularly to reflect changes in the system, model features, and local environmental characteristics (new shops, new living areas,..). For example, if new data sources (e.g., mobile tracking data or other datasets) are added to the model, the infographic should be updated as soon as possible.

2.2.2. Accompanying measures (in addition to the infographic)

In addition to this infographic, a list of considerations/KPI for applying the disregard function (i.e. performing human oversight) should be elaborated. Examples:

- Individual oversight: very different route than recent, previous ones? receiving a similar route for a very/too long time? Route covers fields and/or forests?

- Collective oversight: many disregards? Increase in burglaries?

Finally, a type of complaint mechanism was conceived whereby police officers perform human oversight "by confirmation", meaning that they would have to confirm or accept a route. If they would like to disregard that route, they would have to explain why they would be disregarding the route recommendation. They could motivate this through a predefined list of reasons (urgent matters in another area, proposed route is away from the prioritized area, catching someone in the act,…) which would be complemented by an open box which could be filled in freely (this could then especially be used for improving the system)

- It would be best to use a dynamic infographic through an online platform or dashboard, which allows users to explore data and predictions in real-time based on changing inputs. If leaflets or posters are used, they should include references to dynamic content where more detailed information is available.

2.3. Training course

The second type of measure is a dedicated **training course for police officers or police zones**. This training should at least explain the AI system and how to interpret its output (explainability) and the related infographic. This training should build further on the

detailed information that needs to be provided in instructions for use (IFU) under the AI Act.

- The format of the training can be a combination of in-person training and workshops and online modules/tutorials for flexibility. Regular testing and assessment are of great importance, due to the possible outcomes of the decision-making by the AI model.

- Training components:

    - Introduction to the BigDataPol AI system

        - Explain the goal and the role of the AI system in predicting crime risks and optimising police patrols

        - Explain what types of data/features are used (as mentioned above)

        - Transparency and accountability: feedback options

    - Understanding and interpreting the AI outputs (explainability)

        - Explain the algorithm in a non-technical way and how the AI model uses data to create risk scores and route recommendations

        - Explain AI model limitations: predictions > certainties, bias, the possibility of putting too much trust in historical data + the dynamic nature of prediction and the influence of changing features on the predictions of the AI model

        - How to use the infographic/dashboard to visualize the AI system and understand why certain risks scores are given and routes are recommended

        - How to interpret certain risk scores and what they mean for patrols and crime-mitigating strategies

        - How to interpret why certain routes are recommended and how they are optimized to reduce crime risk

    - Practical application and human oversight

        - Understanding the role of human oversight in the AI model

        - How to juggle between the predictions, decisions, theoretical professional judgment and possible real-time decision-making and situational awareness

        - Explain the disregard and feedback option, how this is processed, when and where this is possible and to what extent, and the importance of providing feedback to optimize the AI model

        - Practical examples and use cases of scenarios where officers are given AI-generated routes and predictions + training of using the disregard option, providing the officers with guidance on how to interpret the output and make decisions so that the 'basic level of reaction' is the same for the whole team

    - Legal and ethical training

        - GDPR principles: how personal data is handled in the AI model, data minimization, privacy by design, …

        - Explain the AI act and how this system is classified, required compliance; transparency, accuracy, fairness, …

        - Explain the limitations of AI models: predictions > certainties, bias, …