



# AI4Citizens

## Use Case Description and Analysis

Leveraging artificial intelligence (AI) for crowd monitoring can help law enforcement officers and related authorities identify threats to public safety more efficiently, support human operators in making quicker, more informed decisions, improve response times and reduce workload. Yet, AI-enhanced surveillance raises critical ethical, societal, and human rights considerations. **The AI4citizens: Responsible AI for Citizen Safety in Future Smart Cities** project, funded by the Research Council of Norway, brings together a multidisciplinary team to explore how an AI system for crowd monitoring and anomaly detection can be designed and deployed responsibly by anonymizing CCTV footage

The AI system under investigation is **PACA: AI-ENHANCED PUBLIC SAFETY MONITORING**, developed in the Centre of AI Research (CAIR), University of Agder, Norway. This AI system is examined exclusively for scientific and research purposes and is not intended for commercial use or real-world deployment. Its dashboard can be accessed at [https://ai4citizens.uia.no/app\\_dashboard](https://ai4citizens.uia.no/app_dashboard). To reduce privacy impact, PACA anonymizes closed-circuit television (CCTV) footage by masking individuals. It then detects anomalies – incidents that may pose a threat to public safety – and generates alerts for human operators in law enforcement monitoring centres. Operators are then meant to assess the alerts and decide whether to dispatch a patrol. This process aims to replace manual video searches by law enforcement officers, thereby shortening response time to incidents, reducing human fatigue, and reducing errors in incident monitoring.

The United Nations Interregional Crime and Justice Institute (UNICRI) Centre for AI and Robotics has contributed to this multidisciplinary effort by examining the responsible AI considerations underpinning this approach, through the associated project **AI4Citizens: Legal, Ethical, and Societal Considerations of Implementing AI Systems for Privacy-Preserving Crowd Monitoring to Improve Public Safety**, in collaboration with BI Norwegian Business School. This publication presents the resulting use case description and analysis conducted by UNICRI.

The analysis presented herein is based on the information available to UNICRI as of June 2025 and reflects discussions held within the project on practical constraints of developing sociotechnical solutions for public safety through AI-enhanced crowd monitoring. It builds upon the foundational [Toolkit for Responsible AI Innovation in Law Enforcement](#) (AI Toolkit), developed by UNICRI and INTERPOL, with funding from the European Union. It includes suggestions presented to the technical team to align the research with [responsible AI innovation principles](#). It does not, however, provide an exhaustive analysis of all human rights and ethical implications associated with this and similar AI systems, nor does it offer generalizable recommendations for all crowd monitoring systems. This work is published in the interest of transparency and knowledge sharing and aims to contribute to ongoing discussions surrounding the responsible design and deployment of AI systems.

## **Responsible AI Innovation Considerations Related to the AI System**

Carrying out AI research, development and deployment responsibly is a complex, multifaceted process that requires involvement from diverse disciplines and areas of expertise. From a legal and ethical perspective, it entails strict adherence to AI ethics principles and respect for human rights law, as well as compliance with the specific legal frameworks applicable in each jurisdiction. As this analysis does not focus on any single jurisdiction, it draws on international human rights law and the [Principles for Responsible AI Innovation in Law Enforcement](#), as outlined by UNICRI and INTERPOL in their AI Toolkit.

In this regard, several overarching considerations must guide the research, development, and deployment of the AI system, especially given the sensitive context of public safety and law enforcement. First, the AI system needs to be used only for legitimate purposes as defined by international human rights law, such as safeguarding public order, public health, national security, and the rights and freedoms of others (see the [Siracusa Principles](#)). In doing so, the requirements of [necessity and proportionality](#) must be complied with. Additionally, the deployment of the AI system must observe the laws of the country where it is deployed. If a specific legal framework governing AI exists, compliance with it must be assessed and ensured. In jurisdictions lacking a dedicated AI regulatory framework, conformity should still be evaluated against other relevant laws, including those related to privacy, data protection, intellectual property and other fundamental rights. At the same time, this includes ensuring that any limitations on rights are permissible under national legislation. In some cases, specific permissions for law enforcement or other public safety agencies to use the AI system may also be required, depending on applicable national laws. Finally, a detailed impact assessment needs to be conducted [regularly throughout the lifecycle of the AI system](#) to assess its effectiveness and the results of its implementation. Whenever the system is updated or upgraded with additional features, these features need to be assessed as well to ensure continued compliance and responsible innovation.

In addition to these overarching considerations, the AI system under review raises specific issues linked to its unique characteristics. The following use case description and analysis aims to explore these issues. It is structured as follows:

- The first column sets the stage by listing various elements of the AI system, related to its [four main components: hardware, software, data and human](#).
- The second column describes the features of the listed elements of the AI system.
- The third column contains additional considerations of the impact of the system from the point of view of human rights and responsible AI innovation.

## Deployers and users

The AI system is primarily meant to be deployed by law enforcement agencies and other entities concerned with public safety (e.g. local governments).

The AI system users are law enforcement/security officers responsible for crowd monitoring within the respective agency or public safety entity.

The legal framework governing the AI system varies depending on the entities deploying it. To ensure the AI Toolkit principle of [Lawfulness](#), the AI system should be used exclusively by public entities that are entrusted with public safety and public order, as these constitute legitimate aims that may support its use under international human rights law.

## Purpose

The AI system aims to enhance law enforcement agencies' ability to detect anomalies (incidents that may constitute threats to public safety), thereby improving public safety while safeguarding privacy and other human rights. As the AI system intends to apply to pre-existing cameras, traditionally operated by officers, but adding an additional layer of privacy protection, it further aims to reduce the impact of surveillance on the rights to privacy and personal data protection.

The AI system needs to be used only for legitimate purposes under international human rights law, such as pursuing public order, public health, national security, and the rights of others ([Siracusa Principles](#)). The political context of the country where the system is deployed is crucial. A democratic context, founded on the Rule of Law, is more likely to use the system for legitimate purposes, thereby reducing interferences with the rights to privacy and non-discrimination (see also the AI Toolkit principles of [Human Autonomy](#) and of [Fairness](#)), and mitigating any chilling effects on other human rights.

The AI system is intended for routine monitoring of public spaces, as opposed to being used only in large events or situations of public emergency.

Routine monitoring increases the risks of indiscriminate surveillance and treatment of all people as potential perpetrators – risks to the right of fair trial and to the presumption of innocence. When considering its deployment, it will be necessary to demonstrate that appropriate mitigation measures are in place to ensure protection of privacy, limitation of data collection and storage, etc.

Implementing the AI system aims to optimize resource allocation – for instance, increasing the speed and efficiency of law enforcement officers, reducing their workload or lowering the number of officers needed for crowd monitoring. These optimization gains should be checked as part of the continuous monitoring of the AI system.

Essential to uphold the AI Toolkit principle of [Efficiency \(Minimization of Harm\)](#), requiring a favourable ratio between the costs and the benefits of using a certain AI system in terms of time, money, human effort, and the impact on the environment.

The AI system is intended to augment, not to fully replace, human officers, who should always oversee and validate the anomaly alerts issued by the AI system before acting.

Essential to uphold the AI Toolkit principles of [human control and oversight \(Human autonomy\)](#). Effective human control and oversight require a well-informed human-in-the-loop, human-on-the-loop and human-in-command. To ensure this, the impact of the AI output on human decision-making needs to be understood, as well as the risks of automation bias perpetuating algorithmic bias (such as racial, gender and other biases).

The AI system includes (i) a video anonymization algorithm (AN) and (ii) an anomaly detection algorithm (AD). They operate in this order:



The AI system may be deployed in three different modalities:

1. **Central system:** The AN and AD models are deployed in central system servers.
  - The CCTV camera sends raw video to the server, where the AN and AD are executed.
2. **Hybrid (Edge-Central system):** The AN model is deployed on the CCTV camera, and the AD on the server.
  - The camera executes the AN and sends the anonymized footage to the server system, where the AD is processed.
3. **Edge system:** Both the AN and AD are deployed and executed in the CCTV camera.
  - AD metadata, such as alert reports, and the anonymized footage are transferred to the server system.

Each modality presents distinct trade-offs in terms of costs, management and scalability. While all modalities interfere with the right to privacy, they offer varying levels of security for personal data. The best approach for data security is the edge system, where no raw footage is transferred, and alerts are detected locally on the camera, followed by the hybrid, where raw data is not transferred from the camera, and the central system, where raw data is transferred from the camera to the server.

From a data protection perspective, non-anonymized personal data should be copied and transferred as little as possible, with storage limited to the shortest feasible period.

Cameras are located in high-traffic public areas such as streets, squares, and transportation hubs, where there is a demonstrable public interest in crowd monitoring.

The location of the cameras and how many are deployed are important metrics to assess necessity and proportionality in reference to the deployment of the AI system.

The right to privacy is still relevant in public spaces. However, the fact that the AI system is deployed only in public spaces lowers the impact on privacy. Video surveillance in public spaces can be justified for the sake of public safety and crime prevention, provided it is done in compliance with the relevant laws.

The number of cameras should be limited to the minimum necessary to achieve the AI system's legitimate purpose. Limiting the number of cameras to the strictly necessary is particularly important for compliance with human rights law, as it will reduce the impact on privacy.

During deployment, AN automatically "anonymizes" all subjects in the footage collected by the CCTV cameras, before any human accesses the data.

It is important to note, as also highlighted by several interviewed experts, that the term "anonymization" is not used in the same sense as in data protection law (such as the [GDPR](#)). In data protection law, "anonymization" requires that the person is not or no longer identifiable. This also includes indirect identification by singling out an individual.

From a legal perspective, terms such as "masking" or "masking algorithm" are more accurate in this context.

AN employs a human figure segmentation algorithm, which was trained on widely used public data sets (such as the [COCO dataset](#)) to detect human figures.

From both an accuracy ([AI Toolkit principle of Minimization of Harm](#)) and a non-discrimination point of view, there would be added value in using more culturally diverse data for the training to develop a system that is more suited to different cultures.

Moreover, the data used for training and testing AI systems should be obtained and used in accordance with laws and regulations of the relevant jurisdiction, including personal data protection and intellectual property law.

After detection of human figures, AN applies full-body masking, which obscures personal attributes such as the face, body, clothes and personal devices of the people depicted in the footage.

- With full-body masking, human figures are detected and then masked with a solid silhouette colour. Device monitors are also detected and masked as they can display personal information.
- The remaining objects, such as bags or backpacks, umbrellas, weapons, etc., are still visible to enhance AD.

The system is also technically able to mask license plate numbers on vehicles. However, this feature was not included for the purposes of this research. From a responsible AI innovation perspective, masking license plate numbers is recommended in case of deployment, as they may be easily linked to identifiable individuals.

AN intends to protect against processing private information for AD and mitigate the potential biases in the AD that might lead to discrimination (or incorrect results).

The mere collection of personal data, as well as its processing by the AI system, interferes with privacy, regardless of whether humans actually access the personal data. However, some experts consider that the fact that humans do not review personal data themselves reduces the impact on privacy.

AN still has a margin of error, originating from two different sources:

As masking is not 100% effective in eliminating all personal identifiers, some personal data is still processed by the AD and reviewed by users. Additional safeguards need to be in place to ensure that the risks to human rights are mitigated.

1. Individuals are not masked because human figures are not correctly detected by the segmentation algorithm; or
2. Individuals are correctly identified and masked, but certain individual characteristics are still visible or can be inferred by looking at the footage (for example, gender features detectable by the silhouette, gait, or whether someone is using a wheelchair).

Ideally, the training dataset should be representative of the context where the AI system will be used and the applicable laws. Developers should understand what data they are using and that the data should be appropriate to the system's purpose.

AD was trained on the [UCF-Crime Dataset](#) and [XD Violence Dataset](#).

From both anomaly detection and non-discrimination points of view, there would be added value in using more culturally diverse data for the training to provide a more adaptable system to different cultures.

AD was trained using a weakly supervised method to learn how to detect normal behaviour – i.e., typical patterns of interactions between humans and/or between humans and objects.

Moreover, the data used for training and testing AI systems should be obtained and used in accordance with laws and regulations of the relevant jurisdiction, including personal data protection and intellectual property law.

- The training data set includes depictions of the following anomalies: abuse, burglary, robbery, stealing, shooting, shoplifting, assault, fighting, arson, explosion, arrest, road accident, and vandalism.

During deployment, AD processes mostly (as there is still a margin of error) anonymized video to detect an anomaly, defined as unusual behaviour which does not conform to normal patterns.

Anomaly detection should stay centred around threats to public security here there is a legitimate public interest for authorities to intervene. However, these need to be aligned with international human rights principles. For example, the system should not flag behaviours which are criminalized in certain countries against the principles of international human rights law (for example, same sex romantic relationships, or women wearing certain clothes). It is also important to assess which crimes are petty (such as theft) and would not pass the necessity and proportionality test in terms of being a threat to public safety and security. Tests should be conducted so that both developers and users can be reassured that the system's anomaly detection does not flag such instances.

AD still has a margin of error. Some things in a culture are considered an anomaly, but not in a different culture.

The anomalies detected correspond to incidents in which there is a legitimate public interest for authorities to intervene.

Regarding the access to the data by the algorithms:

- AN has access to the raw (non-anonymized) footage.
- AD only has access to anonymized footage except in cases where the AN did not fully mask individuals or individual characteristics.

Regarding data storage and access by users, the AI system may be deployed in three different modalities:

**1.** Users can access the anonymized footage (both in real time and post-event).

- Depending on the chosen camera infrastructure and place of deployment, raw footage is stored in the CCTV cameras or the central system servers.
- Depending on each country's laws, raw footage can be accessed by law enforcement or courts subject to a court warrant.

**2.** Users can access the anonymized footage (both in real time and post-event). In this scenario, only anonymized data is stored, and raw data is deleted immediately after anonymization. After AN, the anonymized data cannot be de-anonymized.

**3.** Users can only access the anonymized footage in real time and cannot access the raw footage. No data is stored, and all data is deleted immediately after AN and AD. The only information stored is the generated metadata (timestamp and a textual description of the incident) to provide some context of the situation.

This is important for the system to be considered useful to law enforcement, otherwise it does not serve its purpose. Allowing the officers to review the anonymized footage contributes to increasing the accuracy of the decision-making of the users. The users will make more informed decisions if they can see the anonymized footage.

In case authorities are mandated to keep the raw footage, that needs to be done for a limited period of time and in strict compliance with the law.

For maximum protection of privacy and personal data, authorities cannot routinely access the raw data. However, there may be legitimate reasons to access the data, such as the anomaly resulting in an arrest and criminal case being brought to court. Whether or not raw data needs to be kept and the exact requirements to access the data depend on the country and its laws. Accessing raw footage should, in principle, require approval by a court or other relevant independent authority.

The performance of the AI system is evaluated based on the following metrics:

- AN: whether a (set of) particular personal attribute(s) can still be identified after anonymization, such as face, skin color or gender.
  - Privacy category per attribute: identification of a category of a privacy attribute, such as gender or race.
  - Person identification: whether a person can be identified using querying from a gallery image, matching personal characteristics.
- AD: whether an event is accurately detected as an anomaly or normal.

The performance of the AI system shall be tested with data relevant to the context where the AI system will be deployed.

The fairness of the system depends on several components ([AI Toolkit principle of Fairness](#)), among them:

- The datasets used, the representation of different groups of population (children, minorities, different cultures), the presence or absence of historical biases in the data, etc.
- The training of the model and mitigating possible biases.
- The locations of deployment of the AI system. For example, if it is deployed in privileged areas of the city with many CCTV cameras installed, it might prioritize the safety of the inhabitants of these areas; if it is deployed in disadvantaged areas, it may exacerbate risks of discrimination against the population by treating them as being at a higher risk of becoming potential perpetrators.
- The use of the system, oversight and redress measures.

The system should have a high level of accuracy, in line with the intended purpose and the state-of-the-art of anomaly detection algorithms. Only a system with an appropriate level of accuracy, robustness and cybersecurity can be considered necessary and proportionate to the public interest needs ([AI Toolkit Principle of Minimization of Harm](#)).

While it is not strictly necessary to train the algorithms with data referring to the country where the AI system will be deployed, the accuracy of the AI system needs to be tested for the context in which it will be deployed. This is essential to ensure the [system's robustness – which includes its reliability and security](#) – and reduce the risk of algorithmic bias in the system, which may lead to violations of the right to equality and non-discrimination during police operations based on the system's outputs.

What errors in AI system represent:

- AN: False negative: leaving a person or a device monitor unmasked. False positive: masking something that is not a person or a device monitor.
- AD: False negative: missed or uncaptured anomaly; false positive: false anomaly alert.

AD sensibility threshold is lowered to increase false positives and reduce false negatives so that authorities do not miss possible relevant incidents.

Increasing false positives leads to increased risks to a set of human rights such as privacy, peaceful assembly and presumption of innocence. Increasing false negatives leads to risks to safety and security. Even though decisions to act on the alerts should always be taken by the user, humans are susceptible to biases that affect their interaction with AI outputs. For instance, there is a tendency to defer to the results of automated systems (automation bias). Therefore, for this to be lawful and compliant with responsible AI innovation principles, it is very important that the safeguards around the use are very strict, and users receive proper training to decrease the risk of over-relying on the AI system's outputs.

Bias in the AI system should be tested within the context in which it is deployed. The impact of the AI system on different groups needs to be evaluated. Unfair bias must be duly addressed and mitigated.

The AI system should not differentiate – directly or indirectly (i.e. based on proxies) – protected characteristics (race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth, or other status, or others recognized as such by the law).  
The AI system should not have disproportionately lower performance on certain individuals or groups based on their protected characteristics.

Performance tests, accuracy levels, the composition of the training data as well as any potential disproportionate impacts on different groups are recorded and communicated to the users.

Essential to uphold the AI Toolkit [principle of transparency](#) (Human autonomy).

When AD detects an anomaly, the users receive an alert with contextual information. The region of interest in the footage is highlighted in a different colour.

The user interface should facilitate critical evaluation of the AI system's outputs. This is essential to uphold the AI Toolkit [principles of human control and oversight and human agency \(Human autonomy\)](#). Human review works as a mitigation of the human rights risks emerging from the AI system's limitations.

The information provided when there is an alert is carefully designed to enable informed and impartial decision-making:

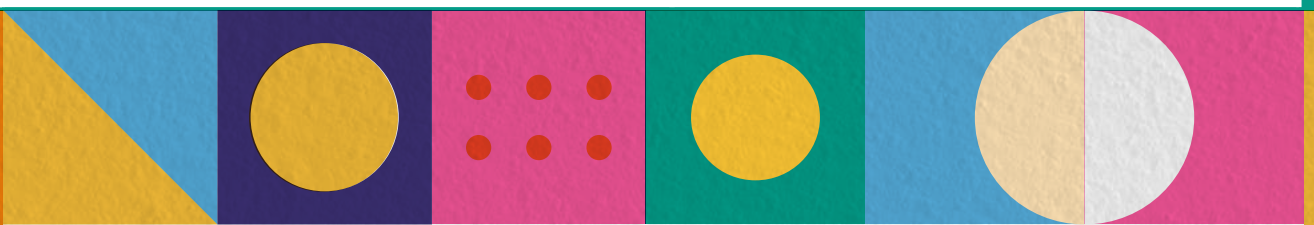
- Anomalies are not labelled as specific crimes, but rather as alerts to potential safety incidents that need further investigation.
- Geolocation, time, and duration of detected anomalies are displayed.
- Detection confidence is displayed on a scale of 0 to 100%.

Labelling certain events as crimes is contrary to the presumption of innocence. Neutral language should be used. Should developers consider using another AI system, such as a Large Language Model, to automatically generate descriptions of the footage, a separate impact assessment should be conducted for this system to assess compliance with ethics and human rights principles and understand data privacy implications.

Users can access the anonymized footage to review the alert and confirm or ignore the anomaly.

The response to the alert (for instance, dispatching a patrol to the location where the anomaly was detected) is based on the above-listed information and the review of the anonymized footage.

Officers using the AI system should review all alerts to understand which cases are false positives that can be ignored and which are true positives that need further action. For adequate human oversight of the system, it is preferable that officers can access the anonymized footage.



## Acknowledgements:

This publication is a product of the **AI4Citizens: Legal, Ethical, and Societal Considerations of Implementing AI Systems for Privacy-Preserving Crowd Monitoring to Improve Public Safety** initiative by the UNICRI and BI Norwegian Business School, with the generous support of the Research Council of Norway. It was written by Inês Gonçalves Ferreira, Maria Eira, and Volha Pashkevich, with participation of Alice Dunglas, and design by Marianna Fassio.

UNICRI wishes to extend its gratitude to all the partners within the project AI4citizens: Responsible AI for Citizen Safety in Future Smart Cities who are running other research tracks and have contributed to the preparation of the report by sharing insights or peer reviewing the draft. Particular thanks and appreciation go to Matilda Dorotic, Emanuela Stagno, Mulugeta Weldezigina Asres, Lei Jiao, and Christian Walter Omlin.

## Additional resources:

On responsible AI innovation in the context of law enforcement: [UNICRI and INTERPOL. \(Revised February 2024\). Toolkit for Responsible AI Innovation in Law Enforcement: \*\*Introduction to Responsible AI Innovation.\*\*](#)

On foundational principles to ensure that AI systems are developed and used for the benefit of society and protect human rights: [UNICRI and INTERPOL. \(Revised February 2024\). Toolkit for Responsible AI Innovation in Law Enforcement: \*\*Principles for Responsible AI Innovation.\*\*](#)

For guidance on how to support law enforcement agencies in implementing and operationalizing responsible AI innovation and documenting decisions throughout the life cycle of an AI system: [UNICRI and INTERPOL. \(Revised February 2024\). Toolkit for Responsible AI Innovation in Law Enforcement: \*\*Responsible AI Innovation in Action Workbook.\*\*](#)

For key concepts and terms in responsible AI innovation: [UNICRI and INTERPOL. \(Revised February 2024\). Toolkit for Responsible AI Innovation in Law Enforcement: \*\*Technical Reference Book.\*\*](#)

For the comprehensive research behind the AI system analyzed in this publication: M. W. Asres, L. Jiao and C. W. Omlin, «Low-Latency Video Anonymization for Crowd Anomaly Detection: Privacy vs. Performance» in IEEE Transactions on Information Forensics and Security, doi: [10.1109/TIFS.2025.3630347.](#)

