

# Best practices to address the legal and ethical implications of AI tech for the security sector

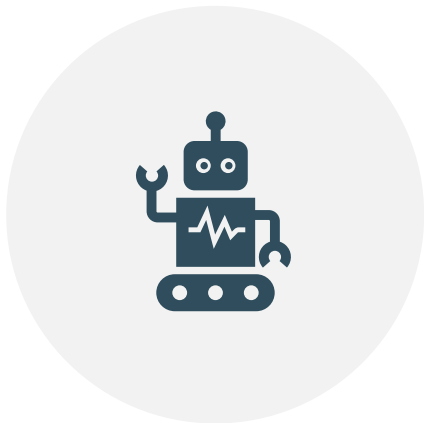
Katherine Quezada Tavárez  
Centre for IT & IP Law (CiTiP) – KU Leuven

AI-Café – 12 May 2022

**Acknowledgment:** Part of the research leading to these results was performed within DARLENE project, funded under EU's Horizon 2020 research and innovation programme under grant agreement No 883297.



# Agenda



AI TECH FOR SECURITY (E.G.  
DARLENE)



OPPORTUNITIES AND RISKS  
OF AI



TOWARDS SOLUTIONS FOR  
ETHICAL AND LEGAL ISSUES

# Augmented Reality

Emerging technology focused on integrating virtual objects into the real-world experience, aligning both the real world and virtual objects with each other in a complementary manner

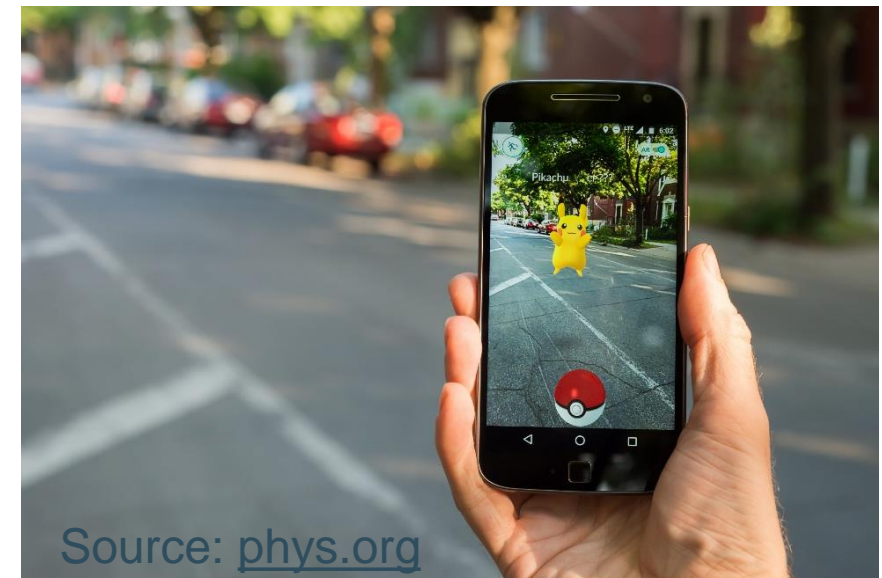
(Azuma, Bailiot, Behringer, Feiner, Julier, MacIntyre, 2001)



Source: [dmexco.com](http://dmexco.com)



Source: [mdc.edu](http://mdc.edu)



Source: [phys.org](http://phys.org)



© Marvel Studios

# AR to improve situational awareness

# Situational awareness

A critical skill in security situations

- LEAs often face **high pressure situations** where forced to **make quick decisions**
- In such circumstances crucial information can be **neglected or delivered too late**
- Rapid / real-time scene interpretation for situations where **time is of the essence**

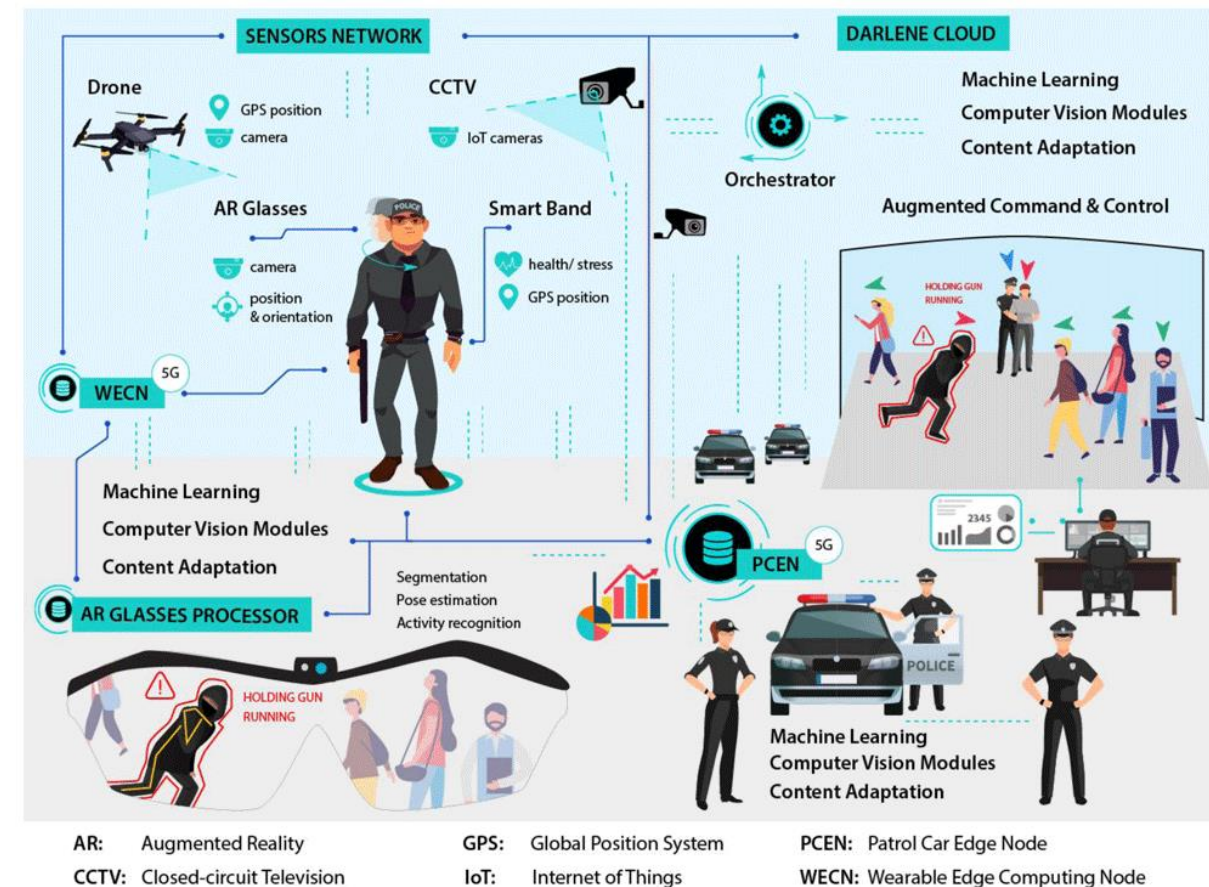


# Cutting-edge AR to

- Enhance officers' situational awareness with
  - Improved optical capabilities
  - Timely reception of crucial information (real-time intelligence)
  - Constant communication with team members



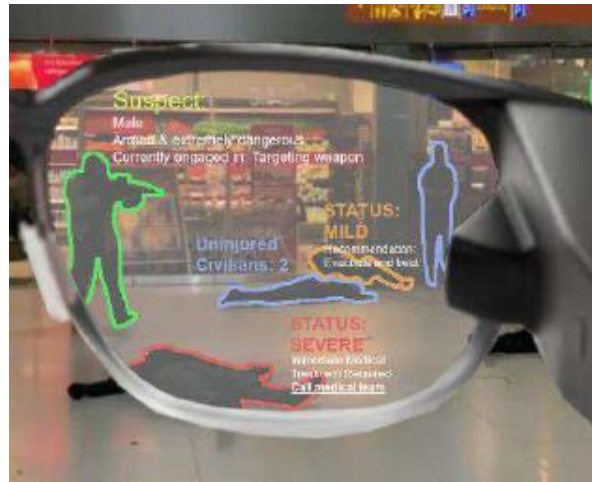
- Rapid processing of the OODA loop (Observe, Orient, Decide and Act)
- Quicker responses to threats
- Better informed tactical decisions



Staying one step ahead  
of adversaries

# Use cases

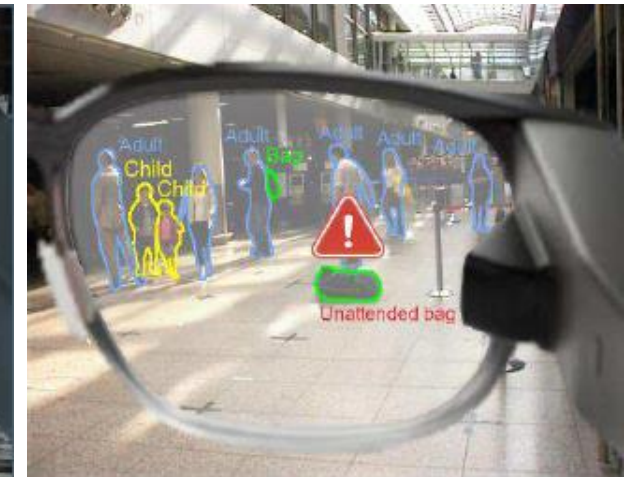
## Rapid visual scene analysis for anomaly detection



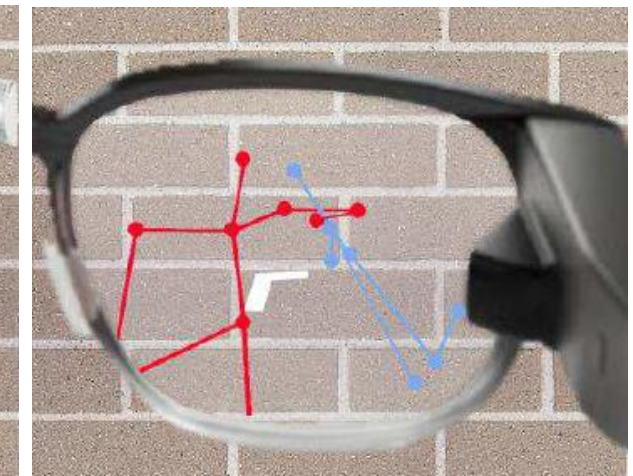
(Crisis management during/after crime)



(Threat detection and neutralization (e.g., potential criminal and explosives))



## Tactical neutralization of human adversaries in the presence of friendlies



(Hostage situation behind the wall)



# AI tech – Opportunities and challenges

## Selected issues(!)

- Automating steps
  - Efficient data processing
  - Cutting costs
  - Enhanced cognitive abilities
  - Reduced human error
- Algorithmic challenges (e.g. the “black box” problem, algorithmic bias)
  - AI governance including data governance
  - Need for human involvement and supervision
  - Need to adapt frameworks and ecosystem

# Towards more legally and ethically sound AI tech for security



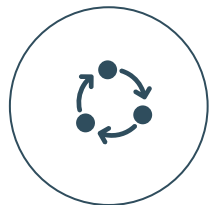
Policy, regulation and guidance



Responsibility by design



Interdisciplinary work



Ongoing law & ethics monitoring



Diversifying teams

# Tech limitations and impacts, e.g. AI evidence

- Evidentiary value?
- Application of exclusionary principle?
- How to ensure reliability (of raw data and its processing)?
- Can we explain how it came into being?
- Meaningful challenging and evaluation in court?



# Overall strategies in security projects (roadmap)

- Extensive analysis of ethical and legal framework (human rights, data protection, criminal procedure, surveillance, evidence, trustworthy AI, AI ethics)
- Guidelines and strategies for system development and implementation (security, transparency, authorization, accountability, compliance)
- Ethical and Legal (Data Protection) Impact Assessment
- Policy feedback

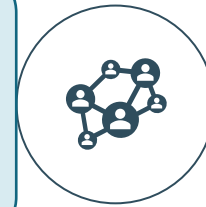
# Towards more legally and ethically sound AI tech for security



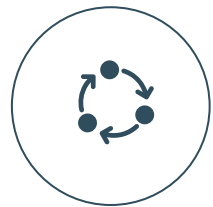
Policy, regulation and guidance



Responsibility by design



Interdisciplinary work



Ongoing law & ethics monitoring



Diversifying teams

# Tech limitations and impacts, e.g. AR

- Compromised device could be fed data to deceive or mislead the user




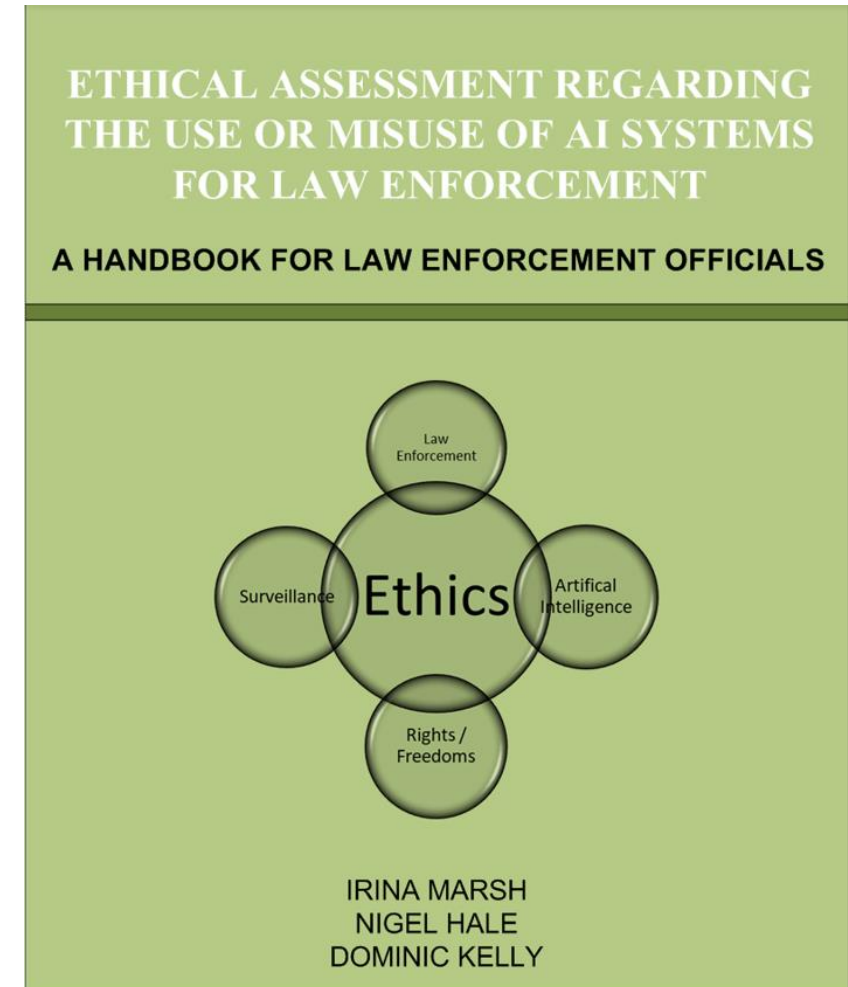
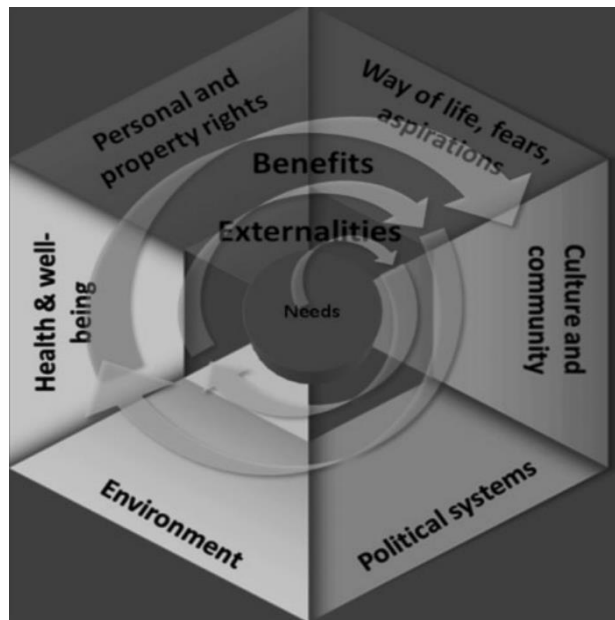
(What the AR device shows)



(What is going on in reality)

# Steering tech development and implementation

 <b>Touchpoint Risk Assessment Table™</b>			
Task no. / title brief task description	Potential ethical, data protection and socio-economic issues	How we propose to address the issues (as discussed with the WP leaders)	Comments of the EAB



# Responsible AI design – Data audit

- Data really (really) matters! (garbage in, garbage out)





# Responsible AI design – Transparent dataset documentation

## Datasheets for Datasets

### Motivation for Dataset Creation

**Why was the dataset created?** (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

**What (other) tasks could the dataset be used for?** Are there obvious tasks for which it should *not* be used?

**Has the dataset been used for any tasks already?** If so, where are the results so others can compare (e.g., links to published papers)?

**Who funded the creation of the dataset?** If there is an associated grant, provide the grant number.

**Any other comments?**

### Dataset Composition

**What are the instances?** (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

**Are relationships between instances made explicit in the data** (e.g., social network links, user/movie ratings, etc.)?

**How many instances of each type are there?**

### Data Collection Process

**How was the data collected?** (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

**Who was involved in the data collection process?** (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

**Over what time-frame was the data collected?** Does the collection time-frame match the creation time-frame?

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

**Does the dataset contain all possible instances?** Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

**If the dataset is a sample, then what is the population?** What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

# Responsible AI design – Transparent model documentation

## Model Card - Smiling Detection in Images

### Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

### Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

### Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

### Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of “fairness” in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

### Training Data

- CelebA [36], training data split.

### Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

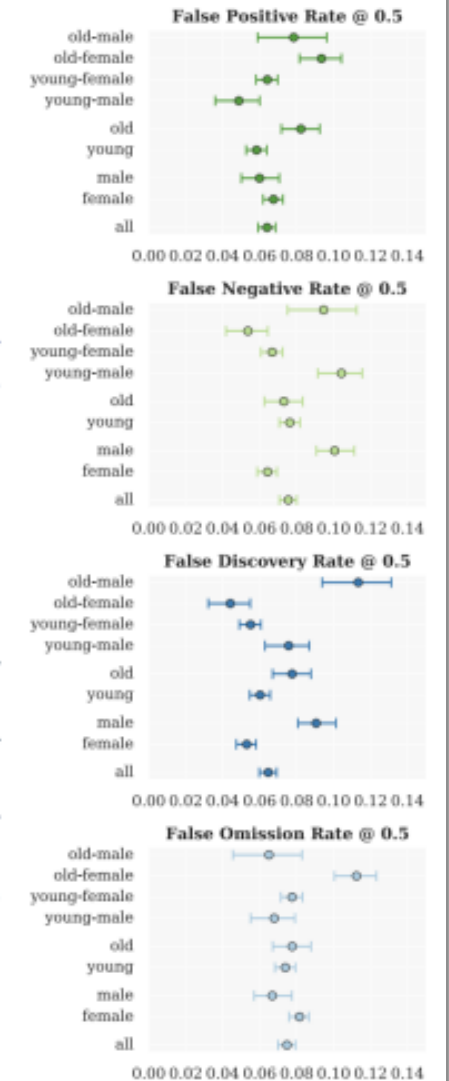
### Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

### Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

### Quantitative Analyses



People with no idea about AI  
saying it will take over the world:

My Neural Network:



AI will take over soon

Source: ifunny.co/

**D**ARLENE

**KU LEUVEN**

**CITIP**

CENTRE FOR IT & IP LAW

# So, don't forget about end-users!



**DARLENE Project** @DarleneProject · May 9



During the first DARLENE training-of-trainers event at the [#HfoeD](#) / [#Police](#) on 3-4 May 2022, the [#prototype](#) of the DARLENE [#AugmentedReality](#) architecture was introduced to and tested by a group of Bavarian [#LEAs](#).

Check out the full article: [darleneproject.eu/darlene-first-...](https://darleneproject.eu/darlene-first-...)



# Towards more legally and ethically sound AI tech for security



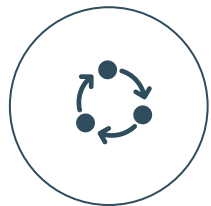
Policy, regulation and guidance



Responsibility by design



Interdisciplinary work



Ongoing law & ethics monitoring



Diversifying teams

# Tech limitations and impacts, e.g. facial recognition



Procedures and the rule of law really (really) matter...



# Multi-pronged approach for interdisciplinary efforts

Mapping

Guidelines

Principles

Requirements



Stakeholder  
engagement,  
e.g. →



# Beware of possible (positive!) collateral effects, e.g.

ACM FAccT Conference

2022 ▾

2021 ▾

2020 ▾

2019 ▾

2018 ▾

Network

Connect

Organization ▾

## Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems

*Ioannis Pastaltzidis, Nikolaos Dimitriou, Katherine Quezada-Tavárez, Stergios Aidinlis, Thomas Marquenie, Agata Gurzawska and Dimitrios Tzovaras*

- Recognizing the risk of using **biased datasets** in real-time detection of crime
- Revealed issues of **overrepresentation of minority subjects** in violence situations that limit the external validity of the dataset for real-time crime detection systems
- Proposed **data augmentation techniques** to rebalance the dataset.



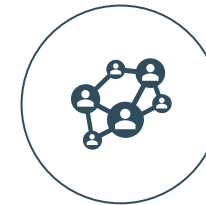
# Towards more legally and ethically sound AI tech for security



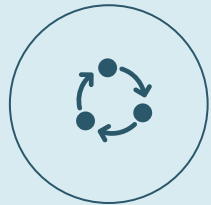
Policy, regulation and guidance



Responsibility by design



Interdisciplinary work



Ongoing law & ethics monitoring



Diversifying teams

# Tech limitations and impacts, e.g. computer vision

Objects Labels Logos Web Properties Safe Search



Screenshot from 2020-04-03 09-51-57.png



Objects Labels Web Properties Safe Search



Screenshot from 2020-04-02 11-51-45.png



# Different technologies, same underlying problem

## Facial recognition use by South Wales Police ruled unlawful

By Jenny Rees  
BBC Wales home affairs correspondent

🕒 11 August 2020

28 July 2021



## Spain: AEPD fines Mercadona €2.5M for illegitimate use of facial recognition system

## Sweden's data watchdog slaps police for unlawful use of Clearview AI

Natasha Lomas @riptari / 11:21 AM GMT+1 • February 12, 2021

🗨 Comment

SECTORAL POLICIES / MIGRATION

Frontex has not complied with all its personal data protection obligations

Brussels, 12/04/2022 (Agence Europe)

On 12 April, the European Data Protection Supervisor (EDPS) reported in a press release that it has reprimanded Frontex for a breach of the data protection regulation applicable to EU institutions, offices, bodies, and agencies.

The institution found that Frontex transferred all of its services "to the cloud without a timely, exhaustive assessment of the data protection risks and without the identification of appropriate mitigating measures or relevant safeguards for processing".

## Frontex has not complied with all its personal data protection obligations

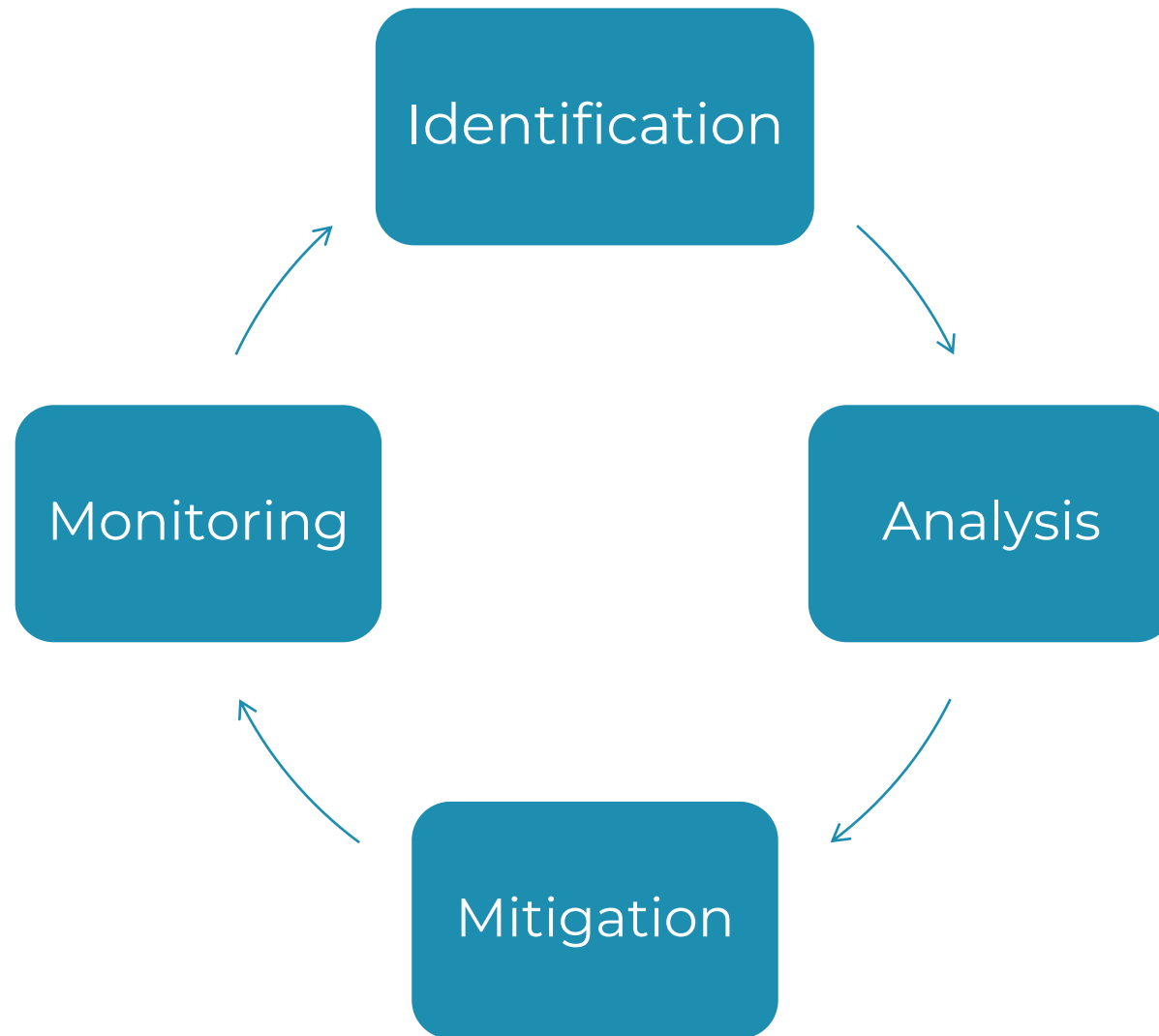
---

*Brussels, 12/04/2022 (Agence Europe)*

On 12 April, the European Data Protection Supervisor (EDPS) reported in a press release that it has reprimanded Frontex for a breach of the data protection regulation applicable to EU institutions, offices, bodies, and agencies.

The institution found that Frontex transferred all of its services “*to the cloud without a timely, exhaustive assessment of the data protection risks and without the identification of appropriate mitigating measures or relevant safeguards for processing*”.

# Holistic approach: E+DPIA & SIA



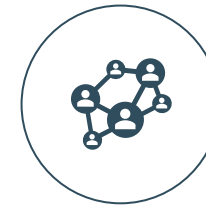
# Towards more legally and ethically sound AI tech for security



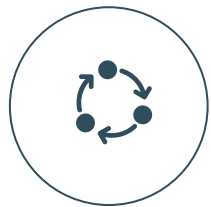
Policy, regulation and guidance



Responsibility by design



Interdisciplinary work



Ongoing law & ethics monitoring



Diversifying teams

# Tech limitations and impacts – not unique to AI

## The deadly truth about a world built for men - from stab vests to car crashes

Crash-test dummies based on the 'average' male are just one example of design that forgets about women - and puts lives at risk



*“Every aspect of science, broadly conceived, is imbued with the characteristics and interests of those who produce it. This does not invalidate every scientific finding as arbitrary or incorrect, but merely points to science’s contingency and reliance on its practitioners—all research and engineering are developed within particular institutions and cultures and with particular problems and purposes in mind.”*

Ben Green (drawing upon Donna Haraway and other feminist scholars)



# Improve team diversity

Particularly in the security sector...



...there's hope on the horizon



# Thank you!

Katherine Quezada Tavárez  
[katherine.quezada@kuleuven.be](mailto:katherine.quezada@kuleuven.be)

KU Leuven Centre for IT & IP Law (CiTiP) - imec  
Sint-Michielsstraat 6, box 3443  
BE-3000 Leuven, Belgium  
<http://www.law.kuleuven.be/citip>

These slides are released under the following Creative Commons Licence:  
Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

# References

- Stergios Aidinlis & Agata Gurzawska (2021). Responsible innovation in Multidisciplinary Research and Innovation Projects. International Society for Professional Innovation Management (ISPIM) proceedings.
- AP4AI, Accountability principles for AI in the internal security domain: [https://www.europol.europa.eu/cms/sites/default/files/documents/Accountability\\_Principles\\_for\\_Artificial\\_Intelligence\\_AP4AI\\_in\\_the\\_Internet\\_Security\\_Domain.pdf](https://www.europol.europa.eu/cms/sites/default/files/documents/Accountability_Principles_for_Artificial_Intelligence_AP4AI_in_the_Internet_Security_Domain.pdf)
- Konstantinos C. Apostolakis, et al. (2021). DARLENE–Improving situational awareness of European law enforcement agents through a combination of augmented reality and artificial intelligence solutions. Open Research Europe, 1(87)
- Ronald Azuma et al. (2001). Recent advances in augmented reality. IEEE computer graphics and applications, 21(6), 34-47
- David Barnard-Wills, Kush Wadhwa, & David Wright (2014). Societal Impact Assessment Manual and Toolkit (D3.1 ASSERT project)
- Donatella Casaburo (2022). AI-Cafe: Law Enforcement AI – Legal and Ethical Challenges of Predictive Policing
- Timnit Gebru, et al. (2018). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

# References

- Ben Green (2021). Data Science as Political Action: Grounding Data Science in a Politics of Justice
- Nicolas Kaiser-Bril (2020). [Google apologizes after its Vision AI produced racist results.](#) Algorithm Watch
- Irina Marsh, Nigel Hale & Dominic Kelly (2021). [Ethical Assessment Regarding the Use or Misuse of AI Systems for Law Enforcement.](#)
- Margaret Mitchell et al. (2019). [Model cards for model reporting.](#) Proceedings of the conference on fairness, accountability, and transparency.
- Ioannis Pastaltzidis et al. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency  
[https://doi.org/10.1145/3531146.3534644.](https://doi.org/10.1145/3531146.3534644)
- Katherine Quezada Tavárez (2021). [Augmented Reality in Law Enforcement from an EU Data Protection Law Perspective.](#) International Review of Penal Law, 92(1), 69-86
- Katherine Quezada-Tavárez, Plixavra Vogiatzoglou, & Sofie Royer (2021). [Legal Challenges in Bringing AI Evidence to the Criminal Courtroom.](#) New Journal of European Criminal Law, 12(4)