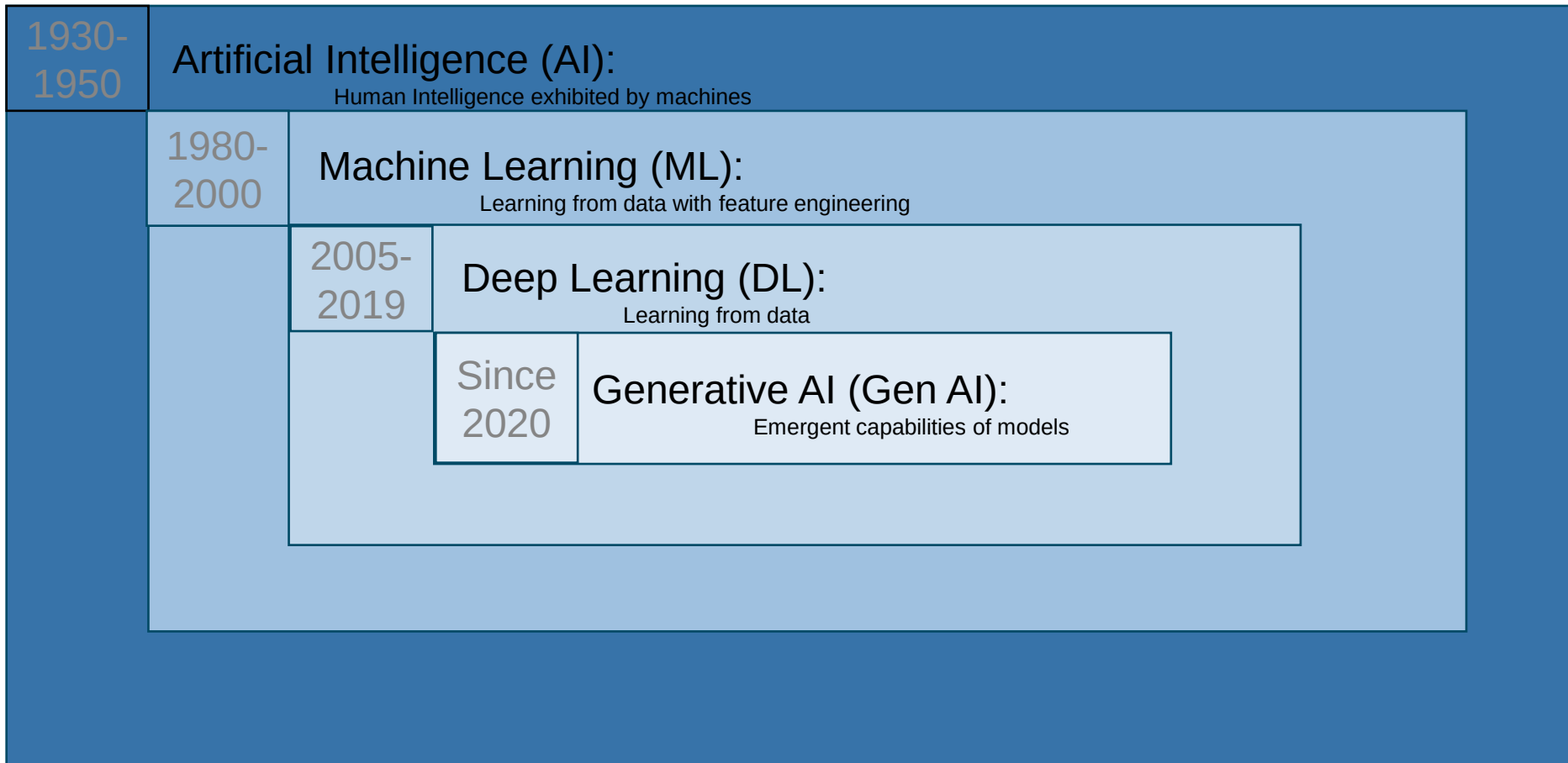


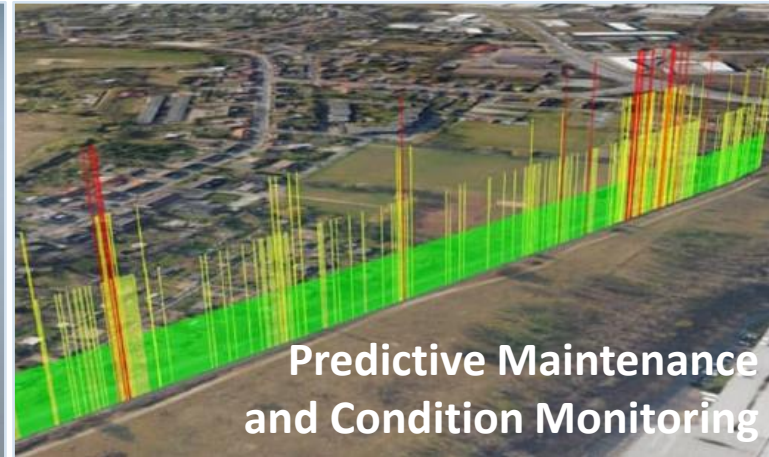
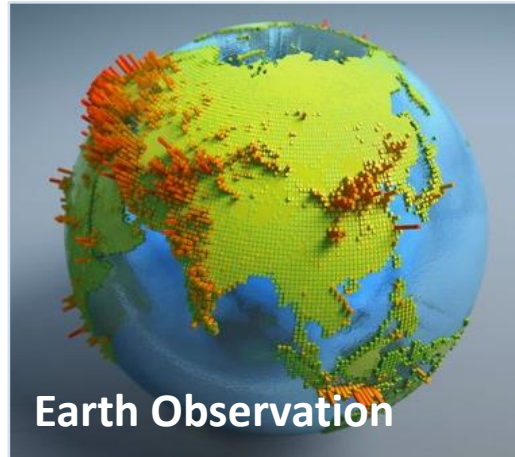
ROBUST AI IN HAZARD PREVENTION

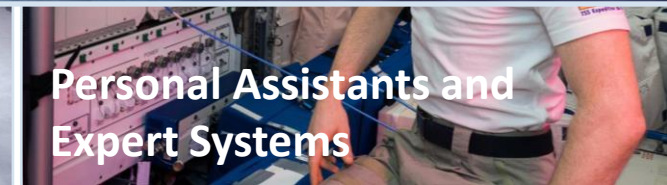
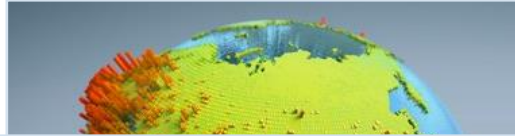
8. Magdeburger Brand- und Explosionsschutztage



AI: A Broader Perspective

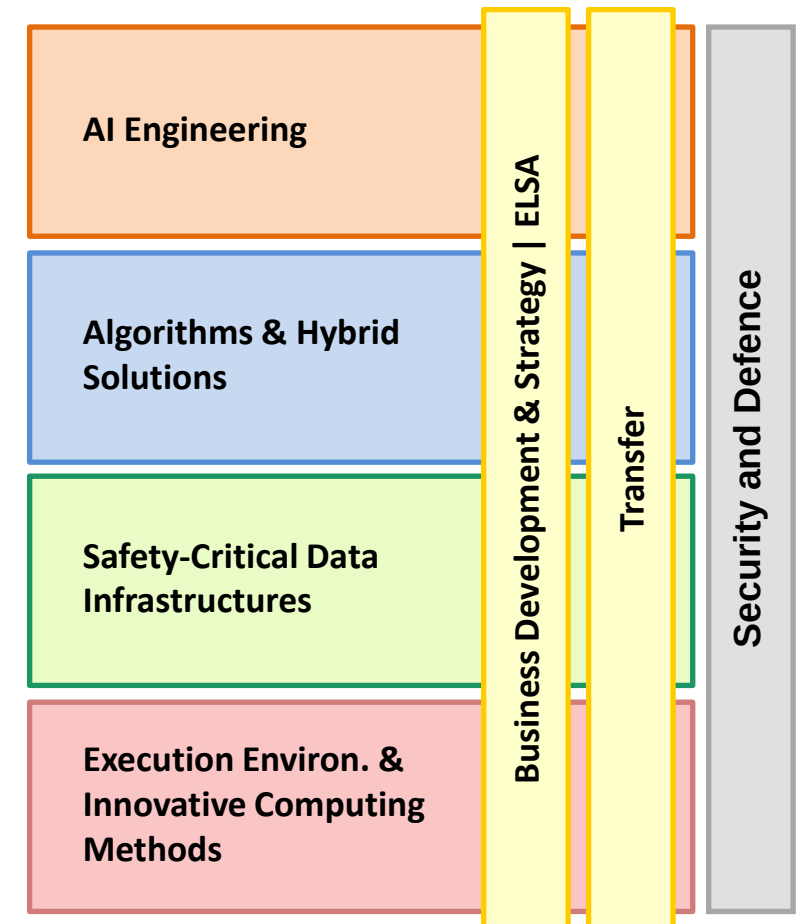




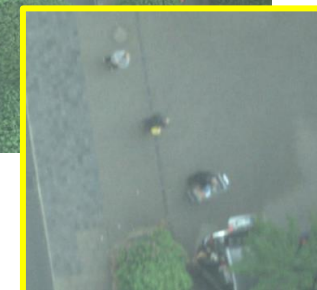
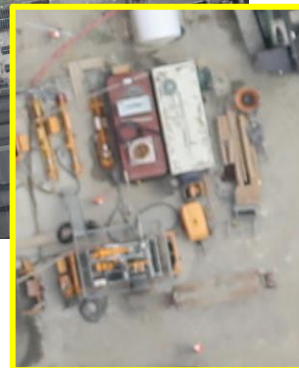
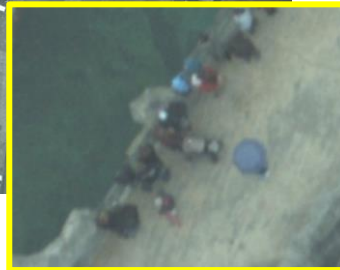


Integrated Management of **Safety & Security** during the **Design | Development & Implementation | Operation | Improvement of AI-based Solutions/Systems** for ambitious/advanced Applications.

- **Safe and Standard-Compliant AI**
 - Guaranteeing safety and security as well as ensuring standard-compliant design of AI algorithms throughout their entire lifecycle.
 - Approval and certification of AI-based systems that are continuously updated or learn.
- **Cyber-Security and -Resilience for AI and Data/Service Ecosystems**
 - Secure management and trustful utilization of sensitive data in open data – e.g. in automotive and rail applications.
 - Protection of data, AI-components and -algorithms, as well as complex AI-based systems against cyberattacks.
- **Autonomous Mobility and Logistics Facilitation**
 - AI-based assistance and automated services to support the introduction and maintenance of advanced mobility and logistics solutions in the market as well as their scalability – e.g. for road, rail, and the maritime domain.



Use Case Example

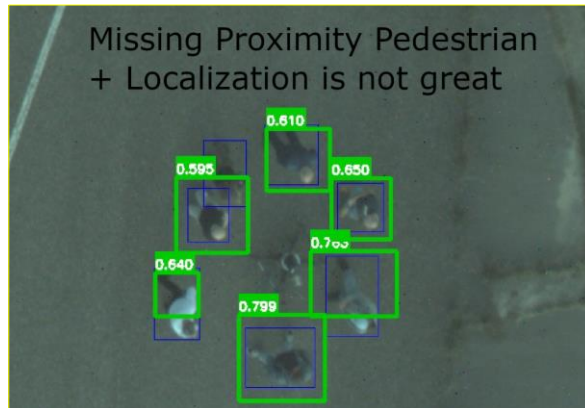


Source:
**SAVING LIVES FROM ABOVE: PERSON DETECTION IN DISASTER RESPONSE
USING DEEP NEURAL NETWORKS**

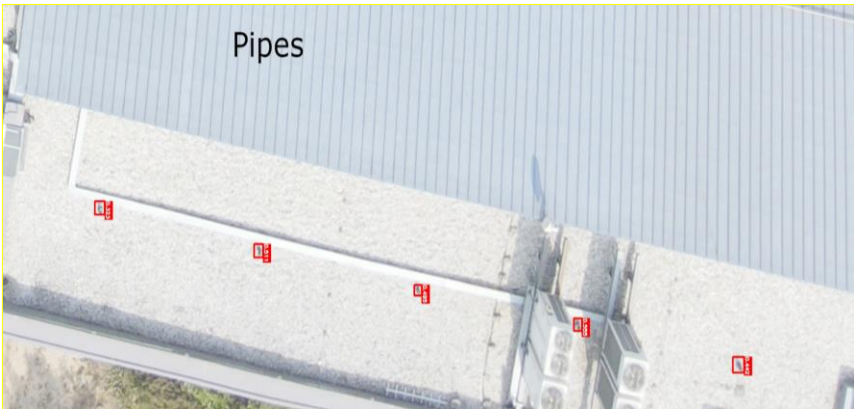
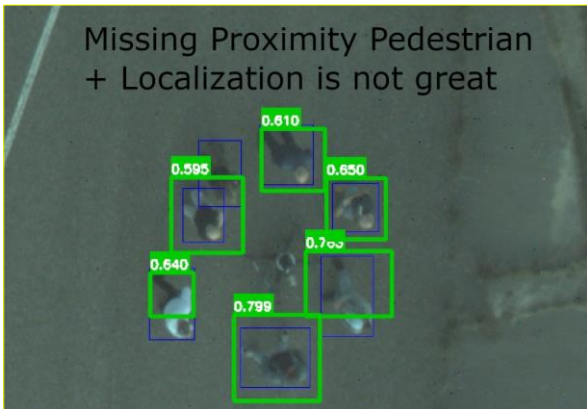
Reza Bahmanyar*, Nina Merkle

Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany
(reza.bahmanyar, nina.merkle)@dlr.de

Use Case Example



Use Case Example



-  wrong detection
-  correct detection

Source: Bahmanyar, R. and Merkle, N.;
<https://doi.org/10.5194/isprs-annals-X-1-W1-2023-343-2023>



Source: xkcd cartoon by Scott Munro (<http://new.pmean.com/machine-learning/>)

Adversarial Examples



88% **tabby cat**

DLR

Adversarial Examples



88% **tabby cat**



99% **guacamole**

Source: "Adversarial Training and Robustness for Multiple Perturbations", Tramèr & Boneh (2019)

Adversarial Examples



88% **tabby cat**

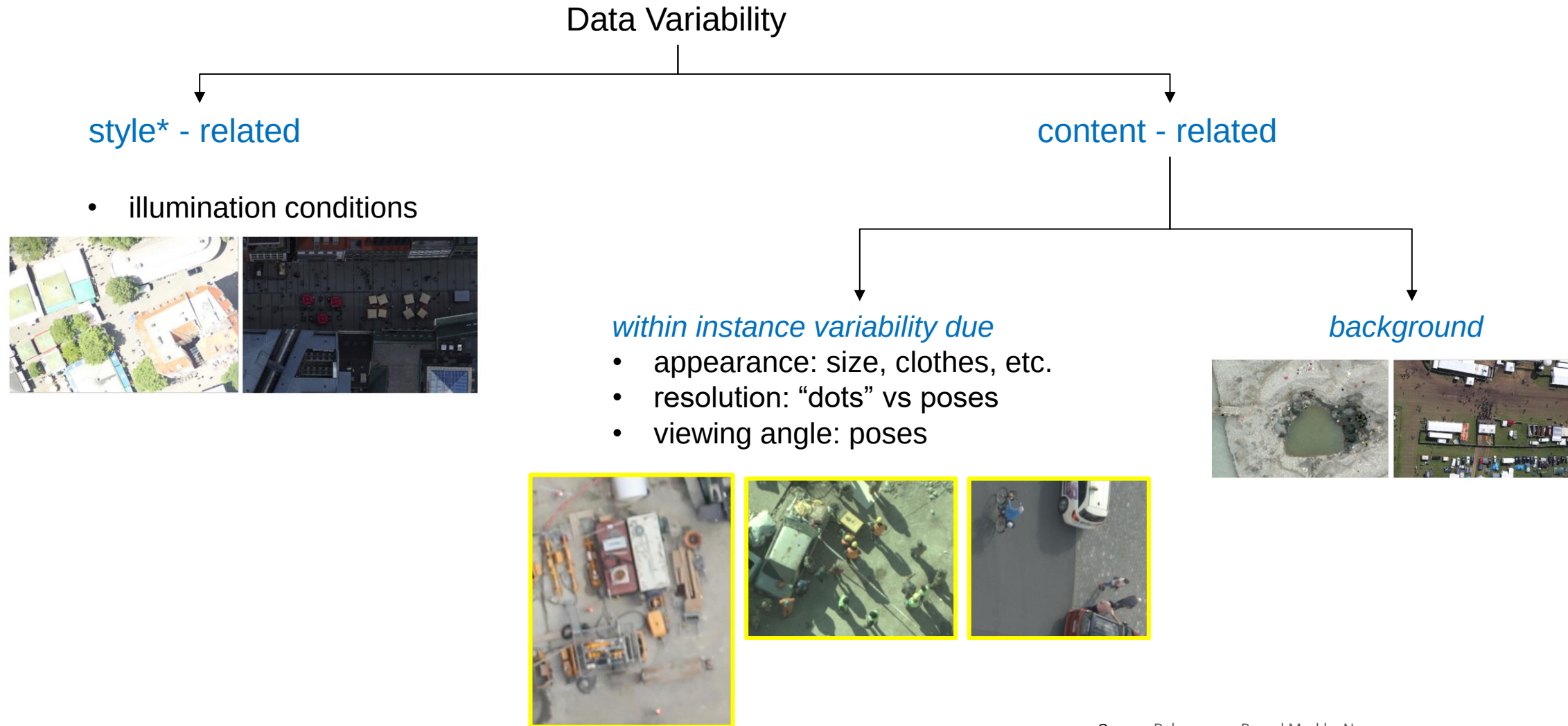


99% **guacamole**

- Robustness
- Explainability

Source: "Adversarial Training and Robustness for Multiple Perturbations", Tramèr & Boneh (2019)

“Adversarial examples” do not need to be malicious:



Source: Bahmanyar, R. and Merkle, N.;
<https://doi.org/10.5194/isprs-annals-X-1-W1-2023-343-2023>

Robustness

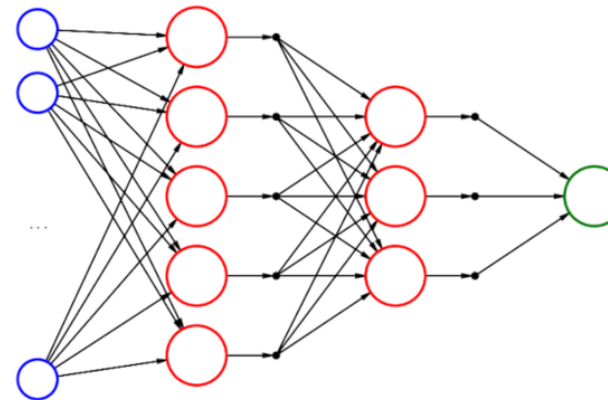
- A models ability to maintain its performance when faced with uncertainties or adversarial conditions

Reasons:

- Noisy data
- Distribution shifts
- Adversarial attacks
- Model complexity

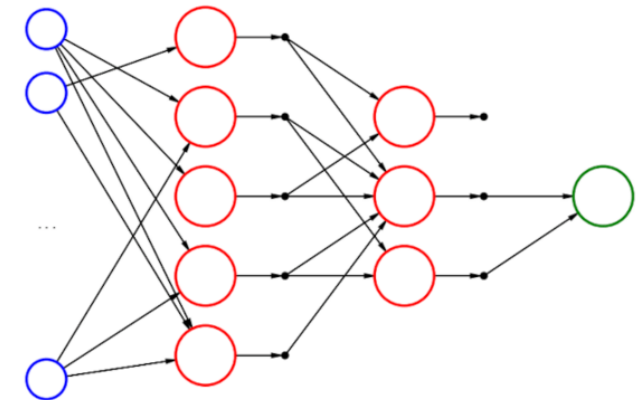
Strategies to allviate:

- Data augmentation
- Regularization
- Ensemble learning
- Transfer learning
- Uncertainty estimation
- Retrival Augmented Generation
- Explainability



Regular Connections

Source: <https://dvgodoy.github.io/dl-visuals/Dropout/>



Applying Dropout

How to increase AI safety

Explainability

- Ensuring that the AI model is not making decisions based on irrelevant or otherwise unfair criteria

Techniques:

- SHAP (SHapley Additive exPlanations)
- LIME (Local interpretable model-agnostic explanations)
- LPR (Layer-wise relevance propagation)
- Saliency map
- Counterfactual (Data) & Mechanistic interpretability (Features)
- Rationalization



Source: <https://www.geeksforgeeks.org/what-is-saliency-map/>

Inputs

- ☐ Image
- ☐ Semantic segmentation
- ☐ Depth map
- ☐ Optical flow



Output

- ☐ Text

Rough example

There is a car 5 meters away in front of the vehicle. The vehicle can reach the car in 5 seconds. {...} Description

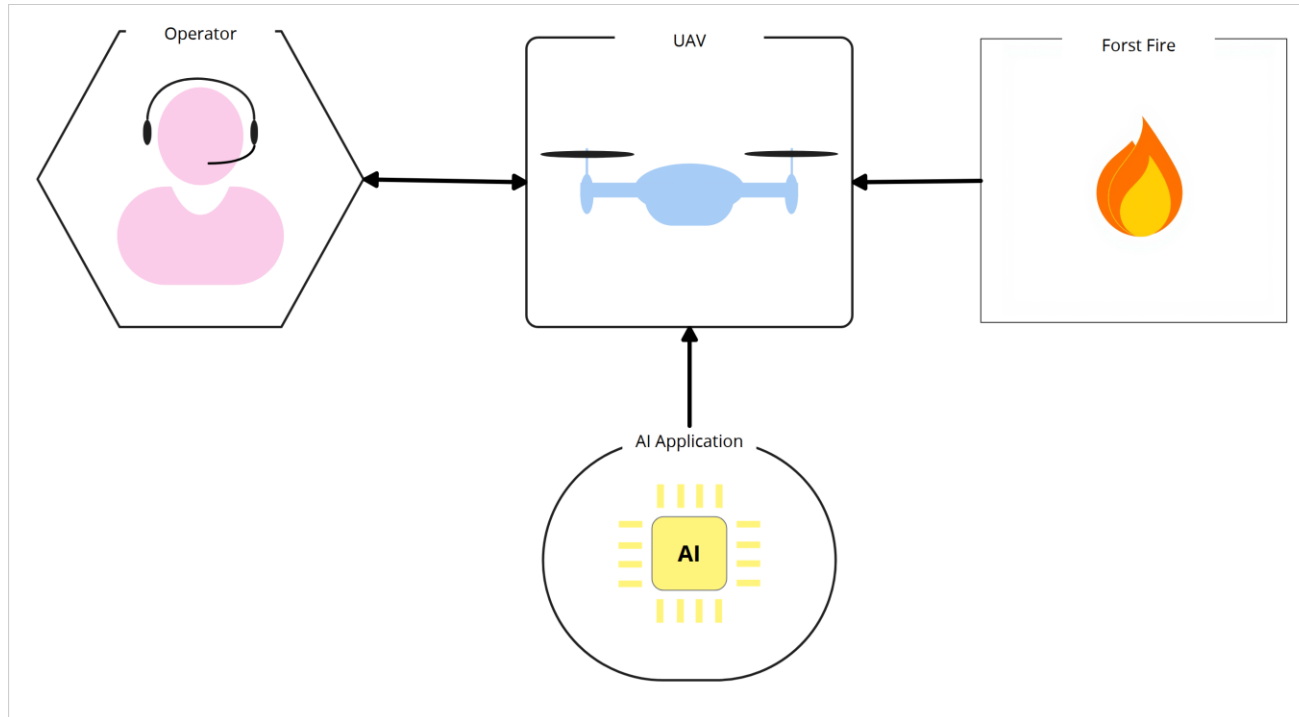
The risk of collision is high. Deduction

Therefore, the vehicle must stop. Action proposition

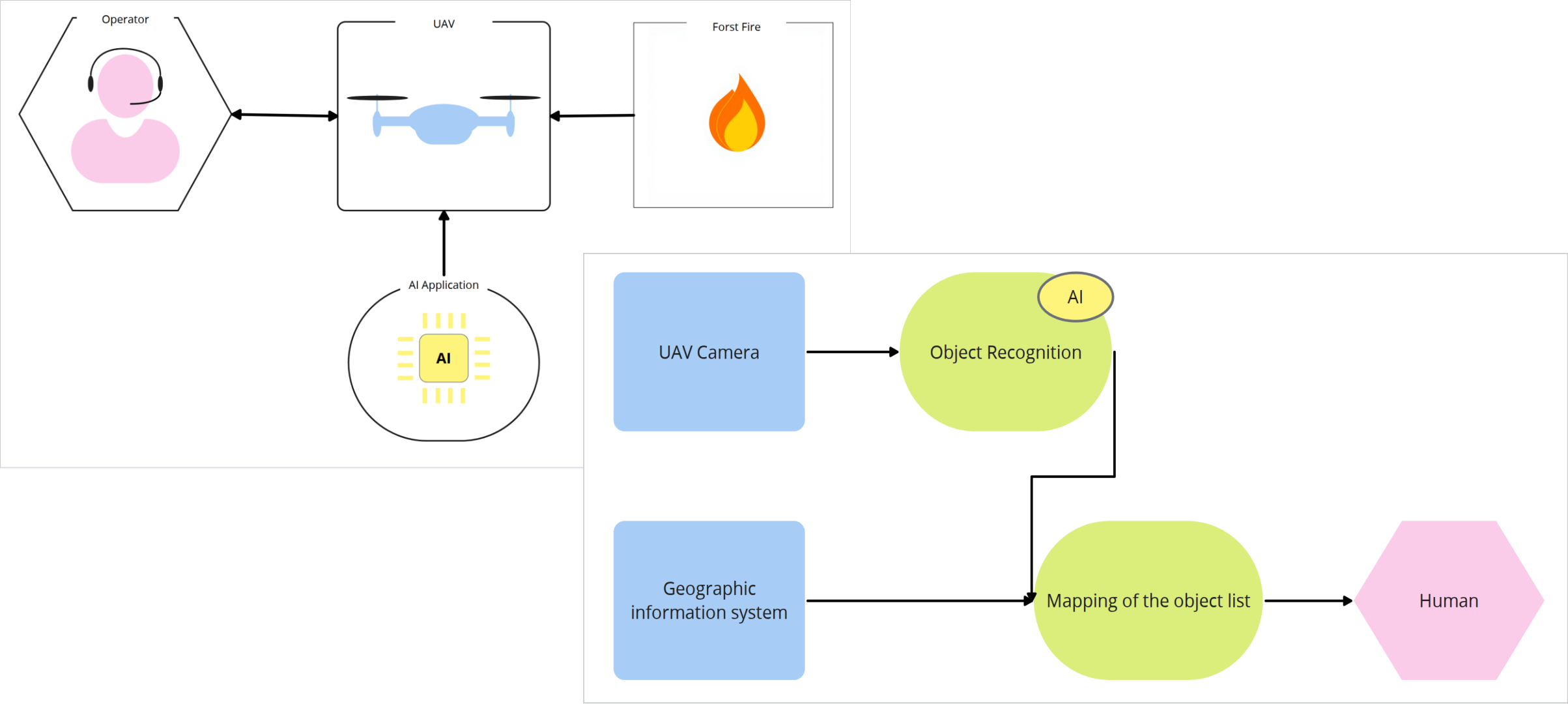
Scene understanding: objects, location, motion, class

Source: Carolyna Hepburn, Project SKIAS (<https://www.dlr.de/en/dw/research-transfer/projects/skias>)

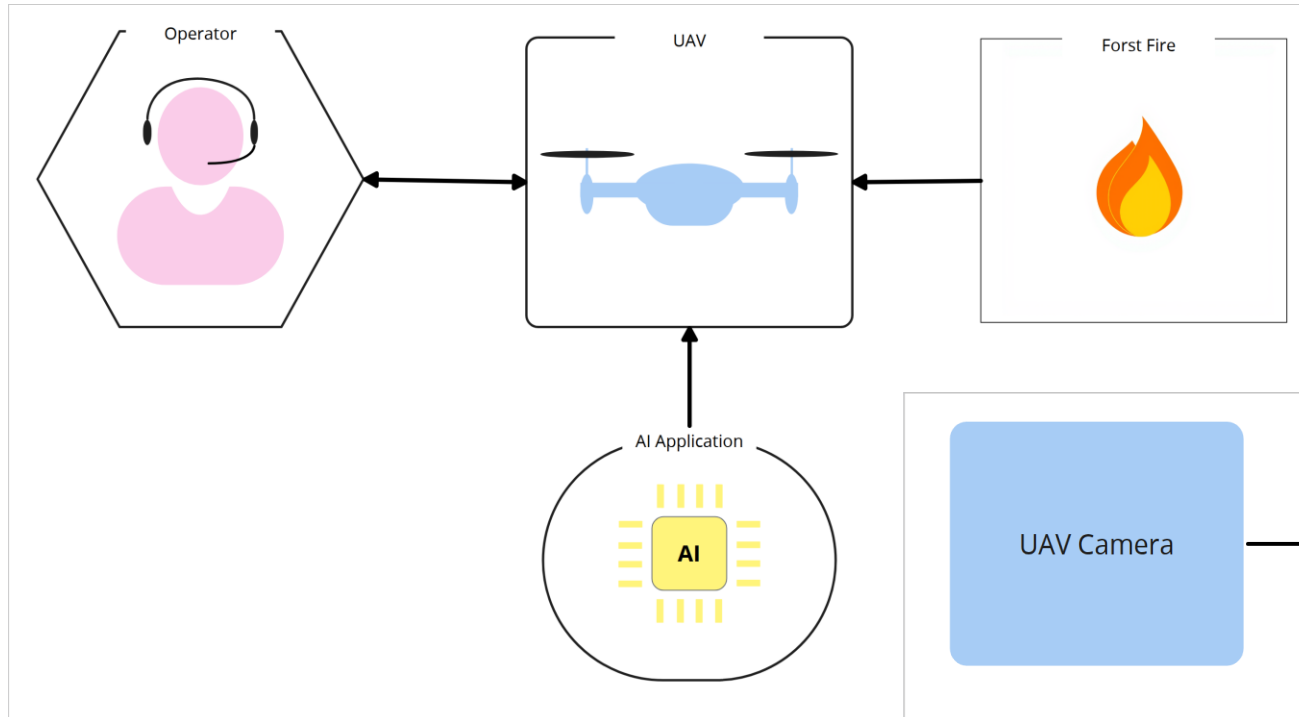
AI in hazard prevention (Example I)



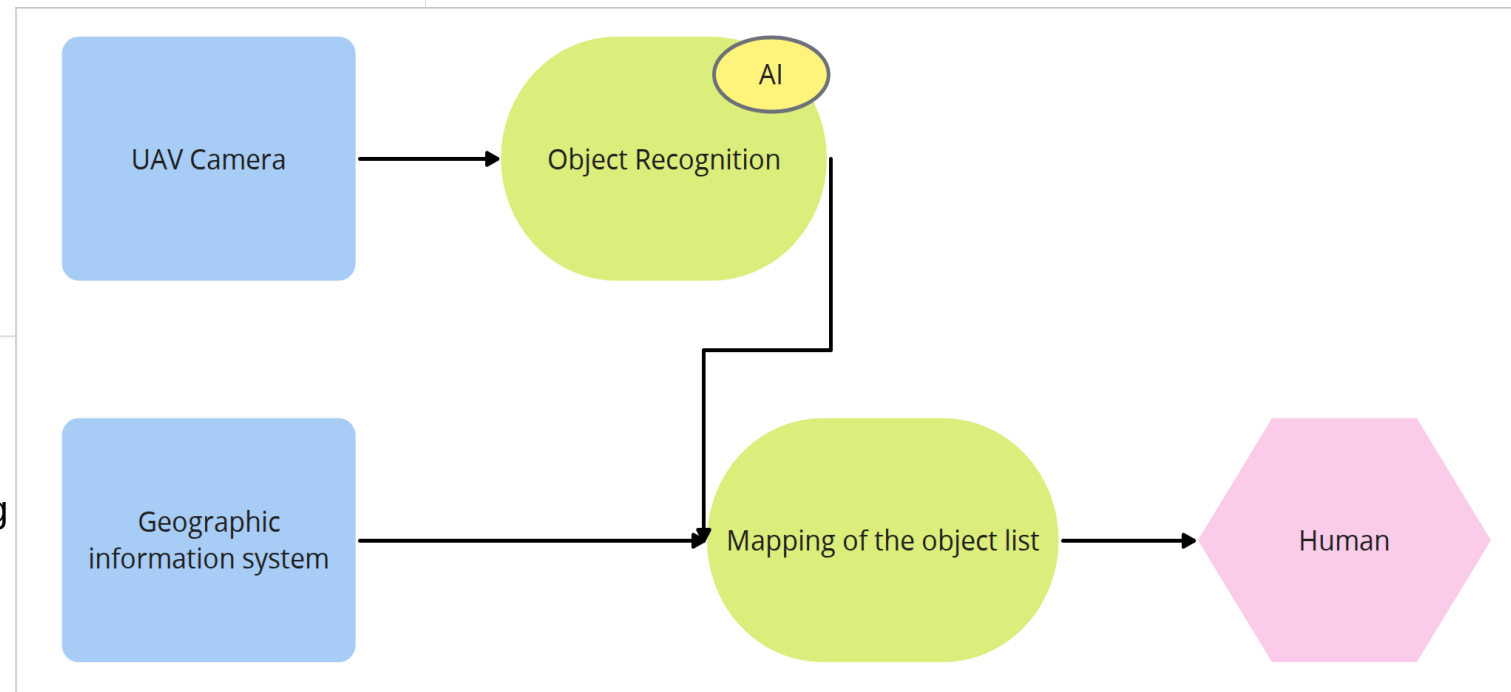
AI in hazard prevention (Example I)



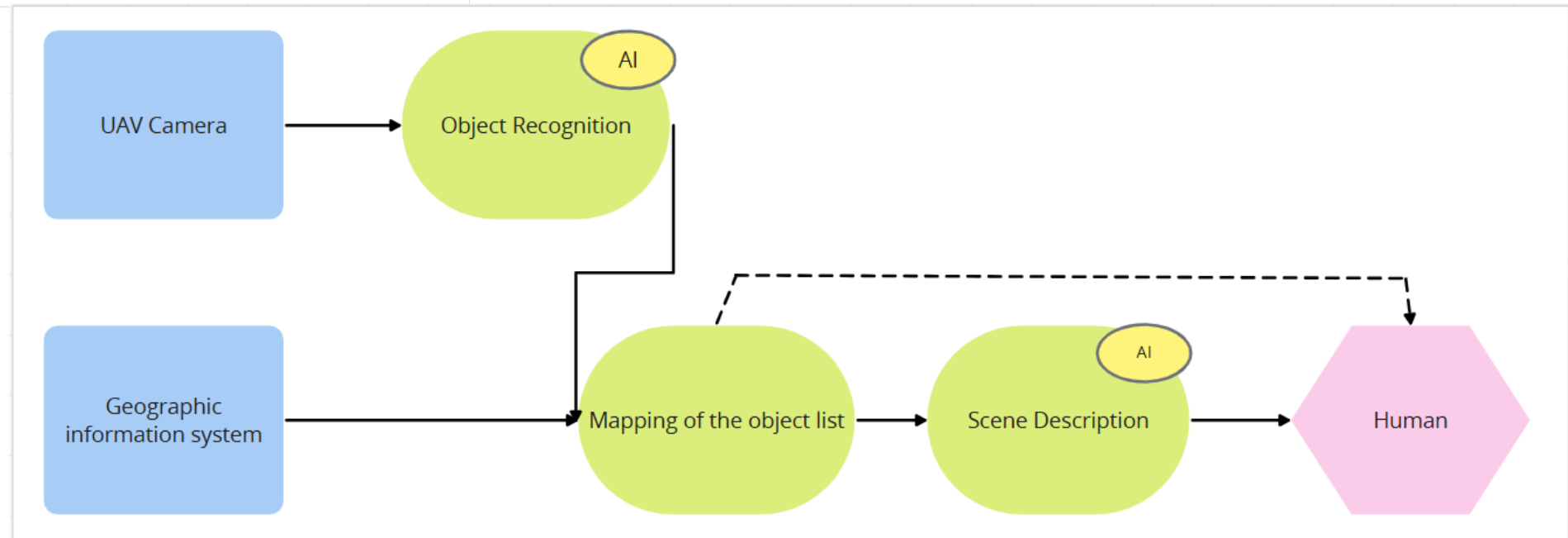
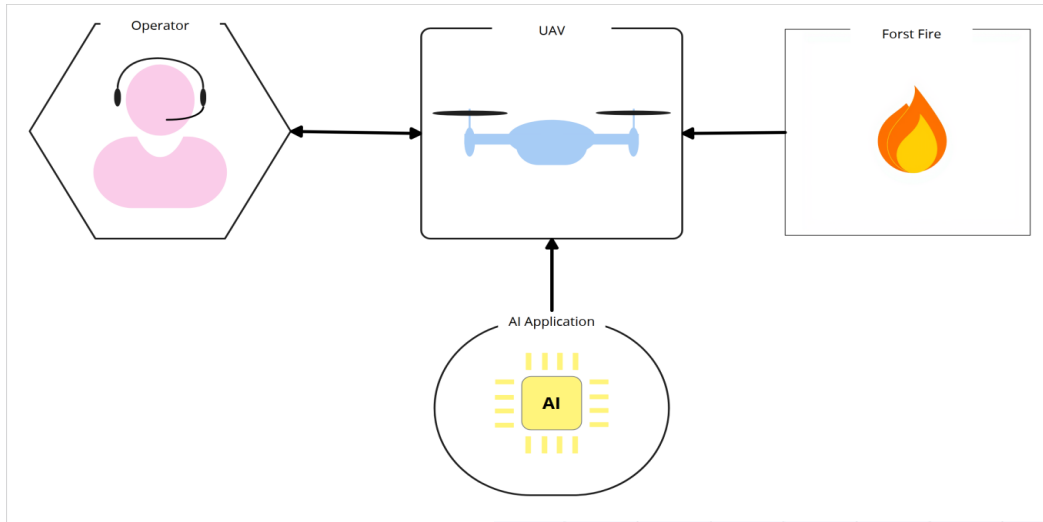
AI in hazard prevention (Example I)



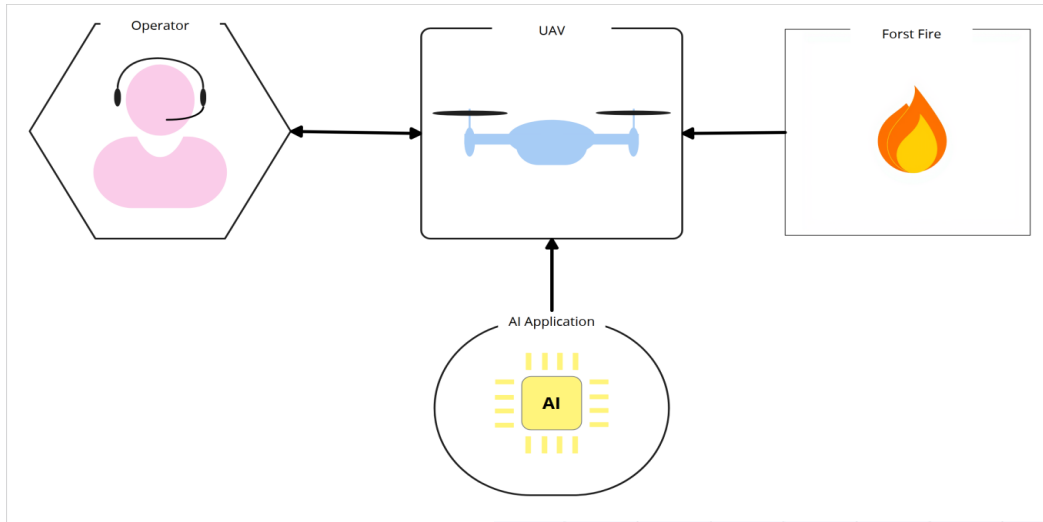
- Operational safety: Not directly effected however AI-processed information might influence decisions
- Information Security: Reliable data source for training and no direct attack on AI components
- Mission aspects: Incorrect coordinates or false detection can make mission more difficult



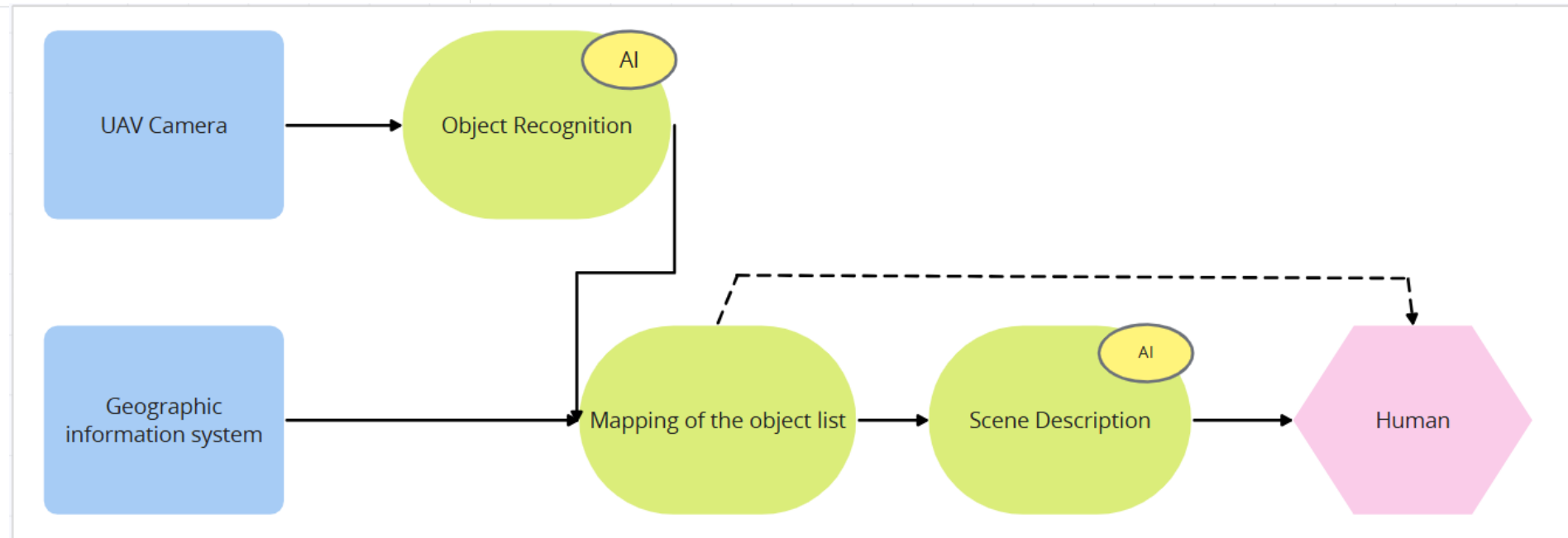
AI in hazard prevention (Example II)



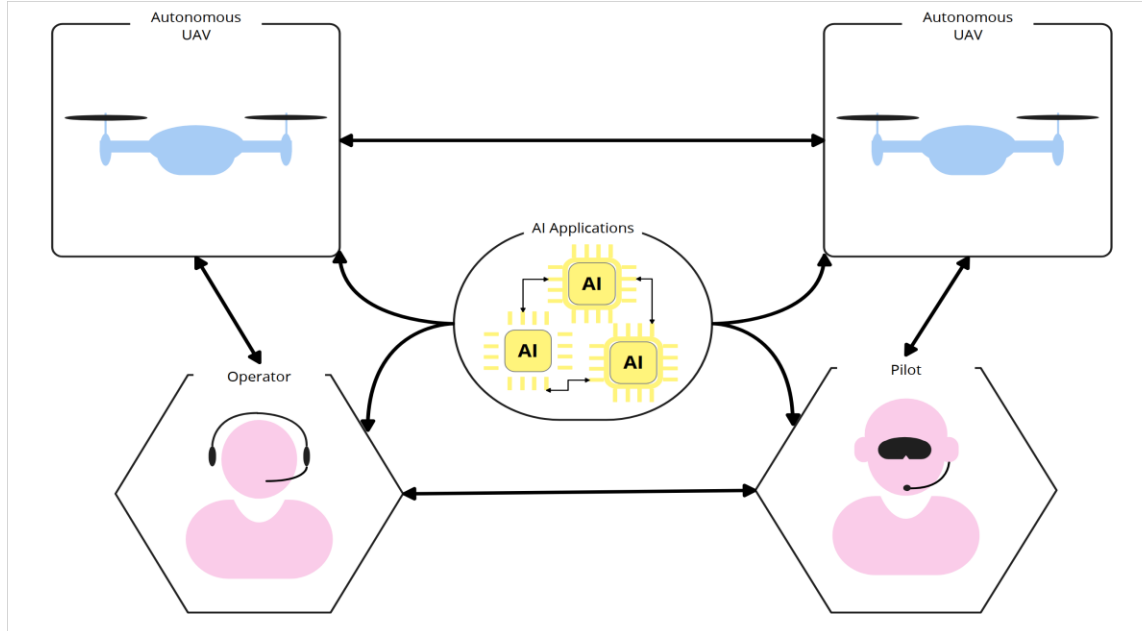
AI in hazard prevention (Example II)



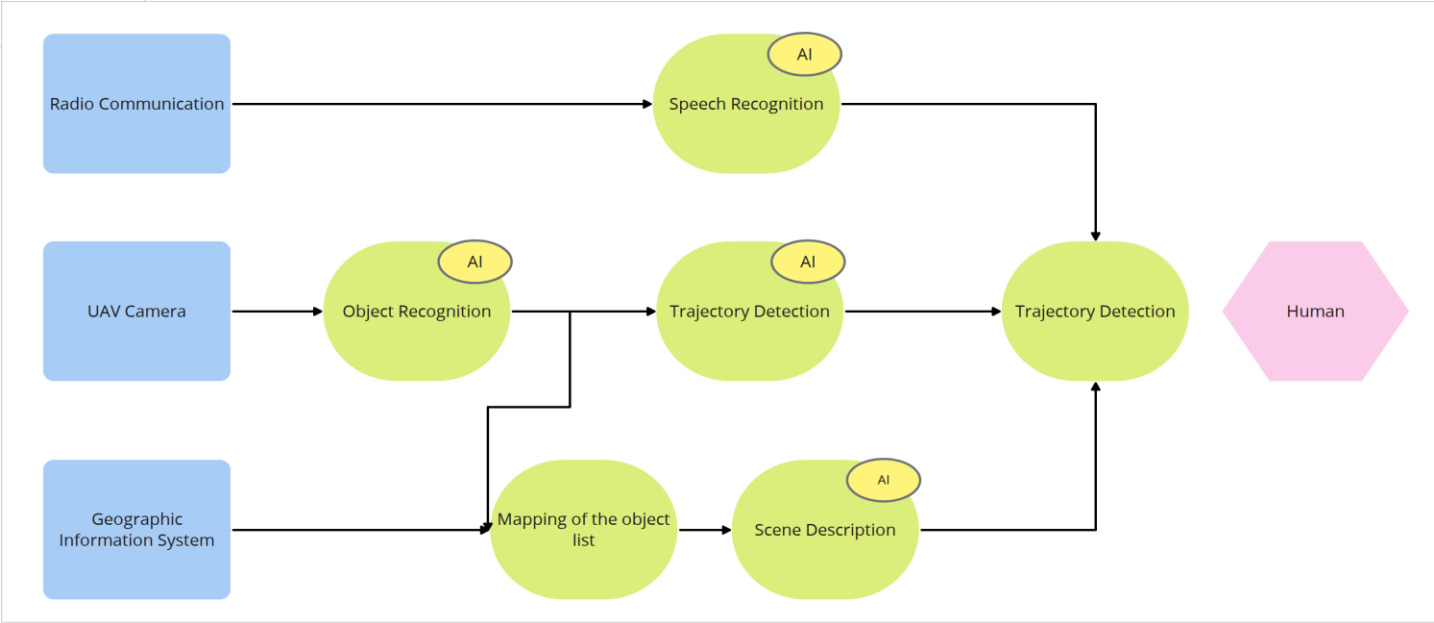
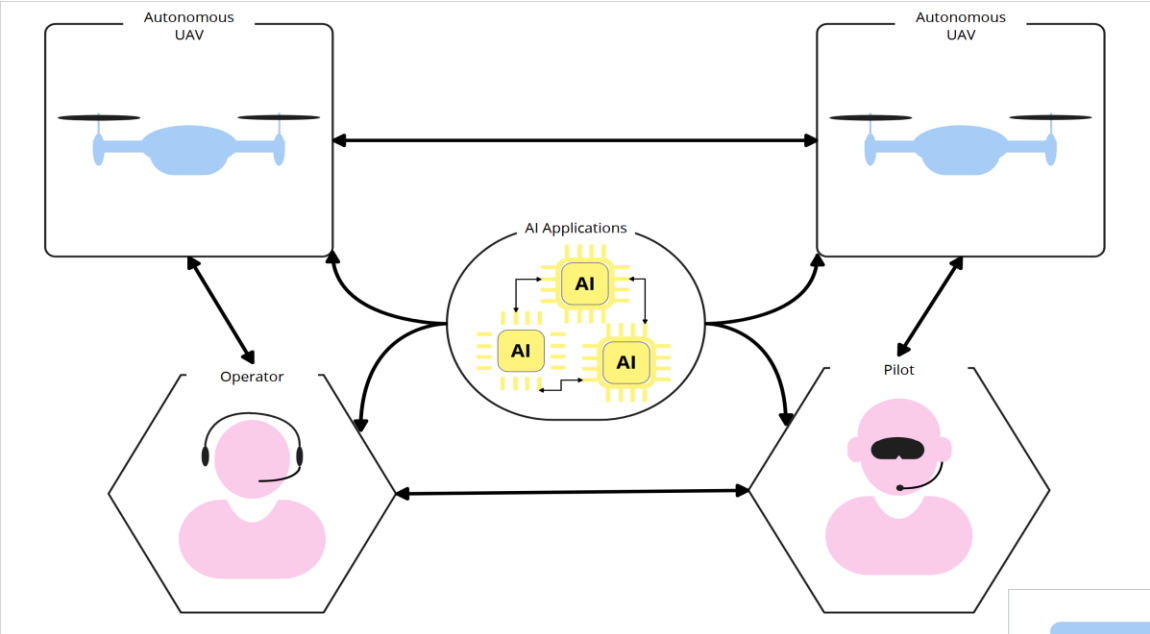
- Operational safety: Failure of AI systems would increase workload of operator
- Information Security: Interlinked AI components make system vulnerable to mistakes; Detection of malfunction/attack is difficult
- Mission aspects: Decisions taken on incorrect information might hinder or prevent mission success



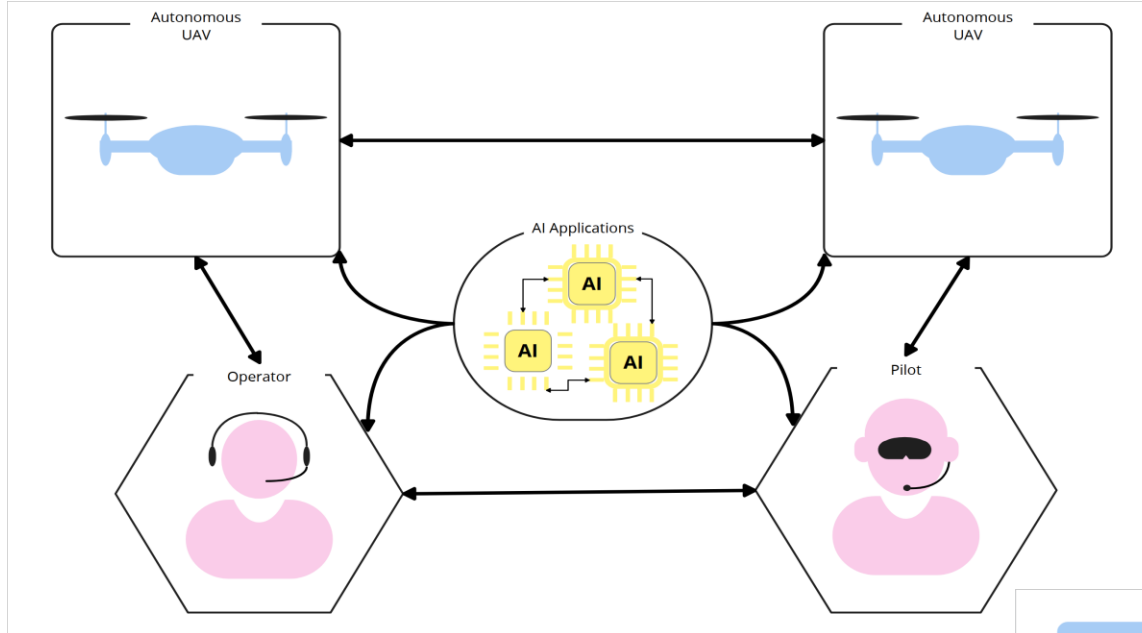
AI in hazard prevention (Example III)



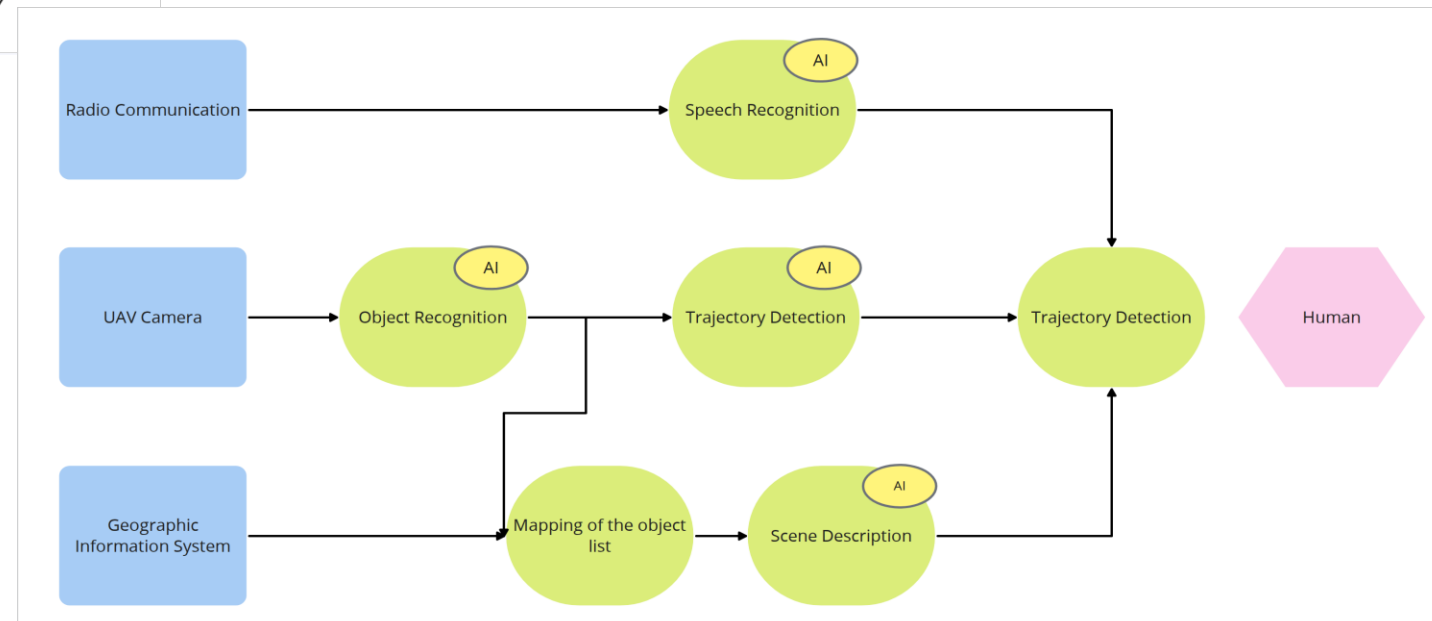
AI in hazard prevention (Example III)



AI in hazard prevention (Example III)



- Operational safety: AI-systems are operating independently and are directly involved in critical decisions
- Information Security: Vulnerabilities in the AI systems have direct impact and can be exploited without much chance of detection
- Mission aspects: Functioning AI-systems are necessary for mission success



Summary



- AI-based systems will increasingly find their way into decision-making processes:
 - reducing the workload of human decision-makers
 - supporting personnel on the ground in the form of autonomous robotic systems.

- AI-based systems will increasingly find their way into decision-making processes:
 - reducing the workload of human decision-makers
 - supporting personnel on the ground in the form of autonomous robotic systems.
- Full integration into safety critical applications needs trustworthy and reliable AI systems:
 - Quality of the data
 - Robustness of interlinked AI components
 - Verified safety argumentation via effect chain analysis

- AI-based systems will increasingly find their way into decision-making processes:
 - reducing the workload of human decision-makers
 - supporting personnel on the ground in the form of autonomous robotic systems.
- Full integration into safety critical applications needs trustworthy and reliable AI systems:
 - Quality of the data
 - Robustness of interlinked AI components
 - Verified safety argumentation via effect chain analysis
- Combination of rule-based, formally verifiable symbolic systems with more flexible and adaptable deep neural networks provides a good basis.

- AI-based systems will increasingly find their way into decision-making processes:
 - reducing the workload of human decision-makers
 - supporting personnel on the ground in the form of autonomous robotic systems.
- Full integration into safety critical applications needs trustworthy and reliable AI systems:
 - Quality of the data
 - Robustness of interlinked AI components
 - Verified safety argumentation via effect chain analysis
- Combination of rule-based, formally verifiable symbolic systems with more flexible and adaptable deep neural networks provides a good basis.
- Enhanced cooperation between research institutions and users of the technology for faster implementation into real life applications.

- AI-based systems will increasingly find their way into decision-making processes:
 - reducing the workload of human decision-makers
 - supporting personnel on the ground in the form of autonomous robotic systems.
- Full integration into safety critical applications needs trustworthy and reliable AI systems:
 - Quality of the data
 - Robustness of interlinked AI components
 - Verified safety argumentation via effect chain analysis
- Combination of rule-based, formally verifiable symbolic systems with more flexible and adaptable deep neural networks provides a good basis.
- Enhanced cooperation between research institutions and users of the technology for faster implementation into real life applications.

Thank you for your attention!



Impressum



Topic: Robust AI in hazard prevention
Date: 2026-03-17 (JJJJ-MM-TT)
Author: Markus Lange
Institute: Institute for AI-Safety and Security
Image sources: All Pictures „DLR (CC BY-NC-ND 4.0)“,
if no other source is stated