Explainable AI at the Edge: Perspectives & Requirements

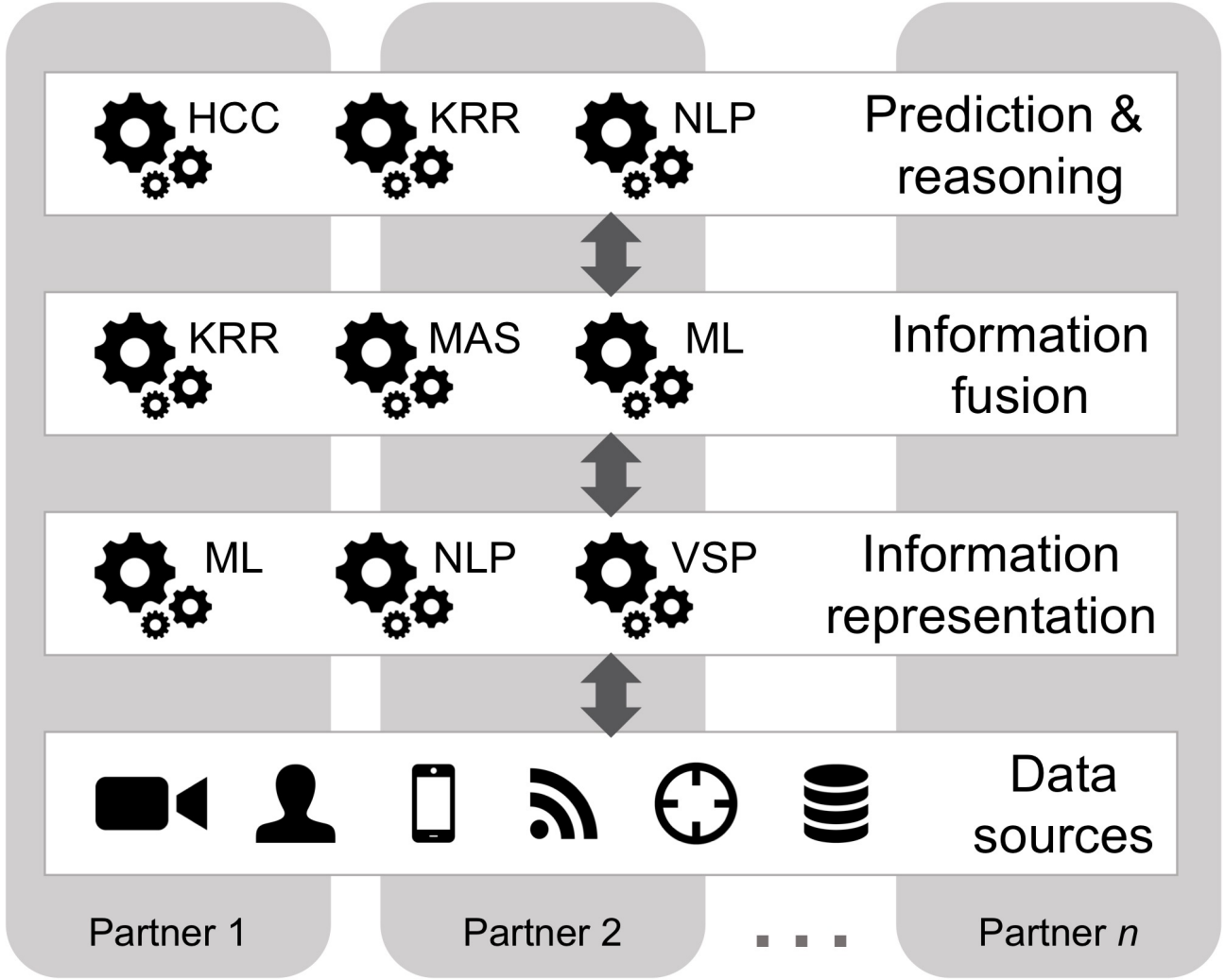
Prof Alun Preece

Cardiff University Crime & Security Research Institute

ARL

[dstl]

Context: Coalition Situational Understanding



"Provide fundamental underpinnings for future coalition distributed analytics and situational understanding in the context of ad-hoc coalition operations at the tactical edge."

Comprehension (insight) = situational awareness + analysis

Understanding (foresight) = comprehension + judgement

HCC human-computer collaboration; **KRR** knowledge representation & reasoning; **NLP** natural language processing; **MAS** multi-agent systems; **ML** machine learning; **VSP** vision & signal processing

ARL

[dstl]

Edge environments

Key features of our edge environments

- rapidly changing situations
- limited access to training data
- noisy, incomplete, uncertain, and erroneous data inputs during operations
- peer adversaries that employ deceptive techniques to defeat algorithms

… are often found in government and public sector applications more generally, as are the joint, interagency, and multinational dimensions.

ARL

[dstl]

Artificial Intelligence (AI) and Intelligence Augmentation (IA)

Challenge: to enable humans and machine agents to operate effectively in joint, interagency, multinational and highly-dispersed teams, in distributed, dynamic, complex, and cluttered environments.

Columbia Pictures, 1963

From the humans' perspective, AI and machine learning are necessary tools to overcome human cognitive limits due to the speed and scale of operations, with the goal of augmenting – not replacing – human cognition and decision-making.

Doug Engelbart

Wikimedia CC BY-SA 3.0

ARL

[dstl]

Human/machine affordances

Consider human and machine assets in terms of their affordances to ISR tasks. Human-machine teaming: each party to exploit the strengths of, and compensate for the weaknesses of, the other.

Machine 

- large-scale data manipulation
- collection/storage of large data volumes
- efficient data movement
- “bias-free” analysis

Human 

- visual/audiolinguistic perception
- sociocultural awareness
- creativity
- broad domain knowledge

ARL

[dstl]

After Crouser & Chang, IEEE TVCG 2012

Perspective 1: the view from GOFAI

Explainability in AI is not a new problem.

In the 1980s AI summer ('Good Old Fashioned AI') it was accepted that explainability was needed for

- system development
- gaining end-user trust

It was also realized that these require different forms of explanation, framed by developers' vs end-users' conceptual models.

The problem wasn't solved before winter arrived!

Perspective 2: the view from stakeholders

Developers are chiefly concerned with building AI applications.

Theorists are chiefly concerned with understanding and advancing AI.

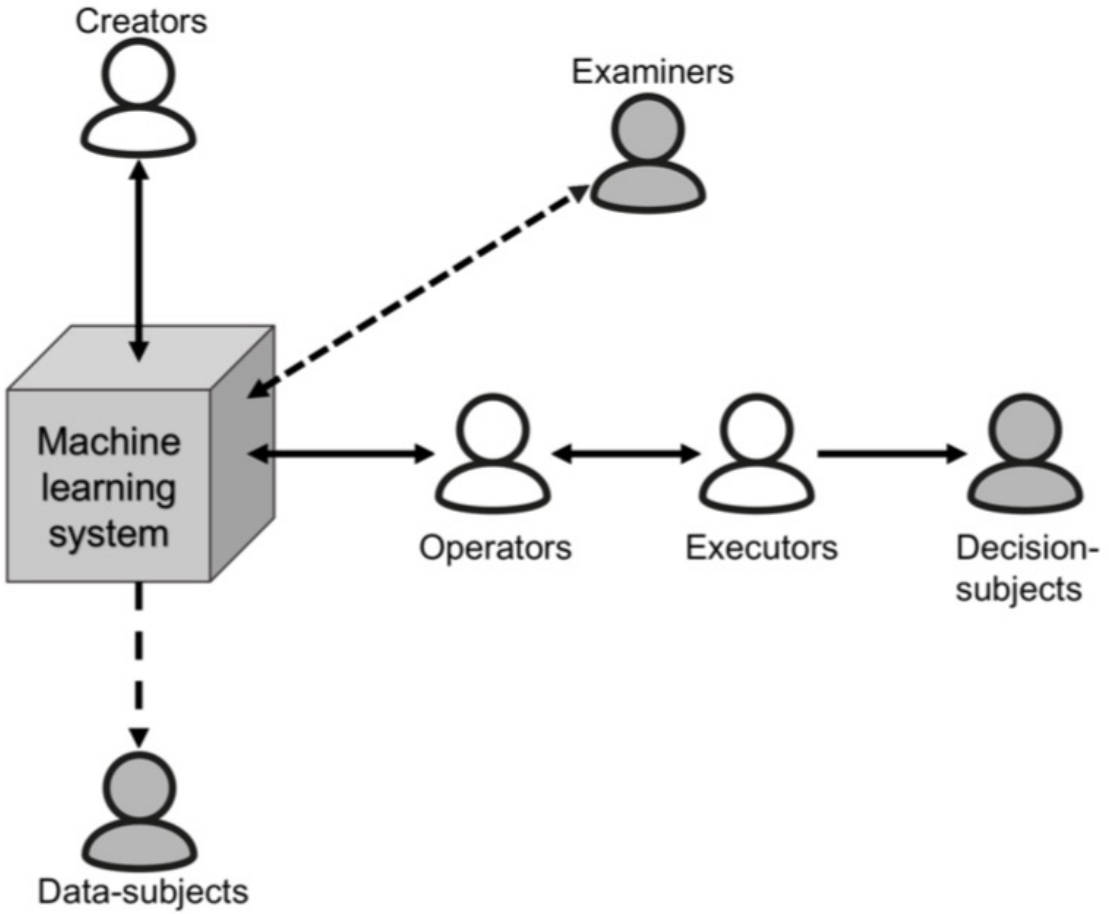
Ethicists are chiefly concerned with fairness, accountability, transparency and related societal aspects of AI.

Users are chiefly concerned with using AI systems.

The first three of these communities are well-represented in the AI interpretability literature. The fourth will ultimately determine how long summer lasts.

"Stakeholders in Explainable AI" *AAAI FSS* 2019
http://sl.dais-ita.org/science-library/paper/doc-3010

Perspective 3:
the view from explanation recipients



"Interpretable to Whom?" ICML *WHI* 2018
https://arxiv.org/abs/1806.07552

Argues that a machine learning system's interpretability should be defined in relation to a specific agent or task: we should not ask if the system is interpretable, but to whom is it interpretable.

ARL

[dstl]

Takeaway: explanation is (just) a software requirement

Explanation requirements are recipient-dependent and need to be **elicited** and **specified** at the outset of an AI project.

Usually have **functional** and **non-functional** aspects.

Need to be **tested**.

Will often relate to **verification**, **validation** and **assurance**:

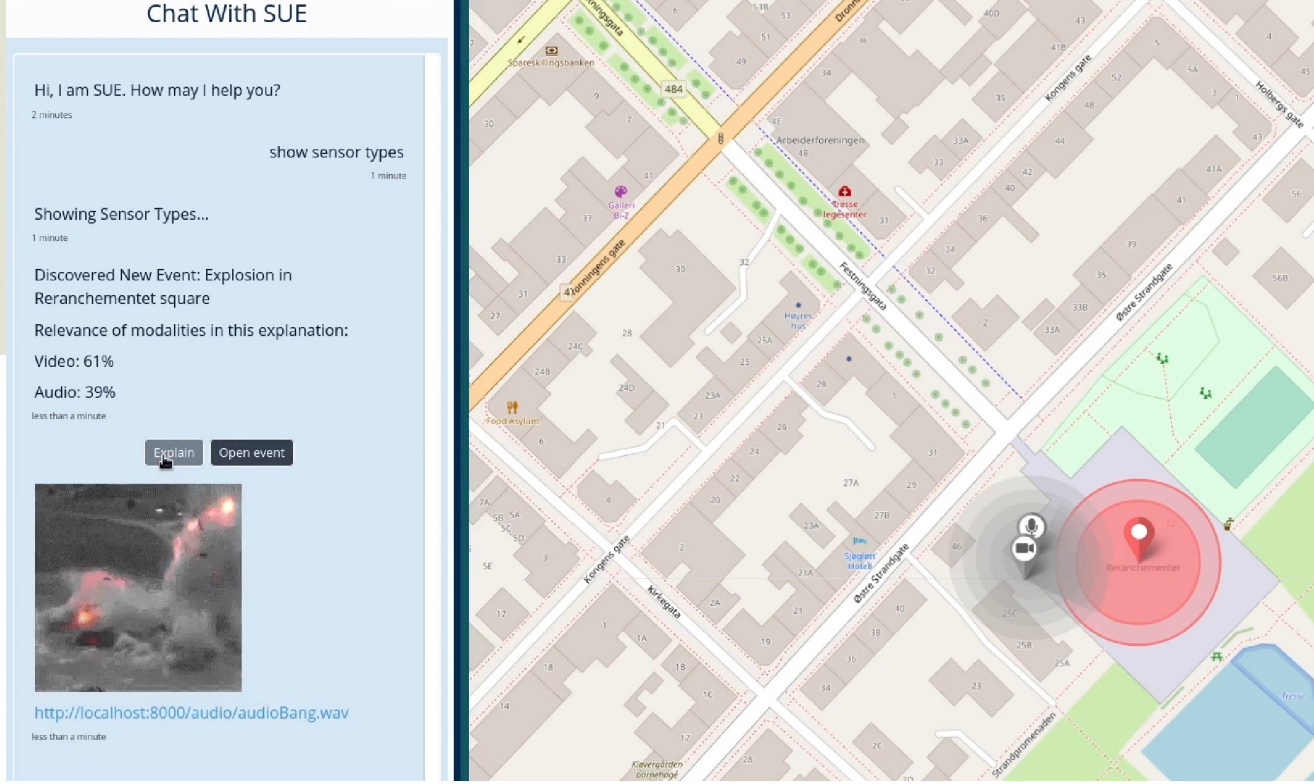
- ✅ verification **shows that the AI system "did the thing right"**
- ✅ validation **shows that the AI system "did the right thing"**
- ✅ assurance **shows that the AI system "does the right thing right"**

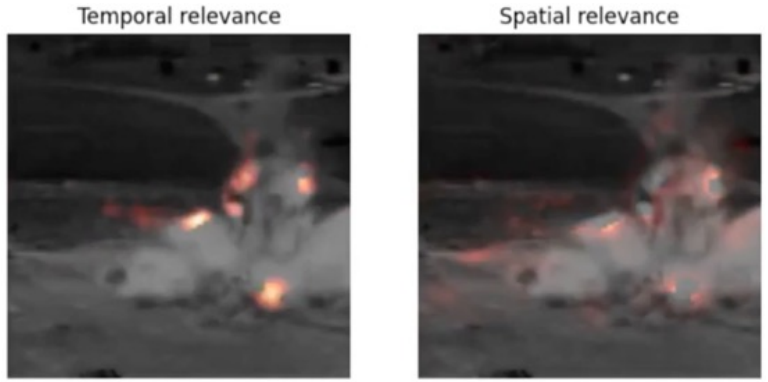
ARL

[dstl]

Example: rapid trust calibration

“Long-term [user] interaction would not be necessary for an AI system with the properties of interpretability and uncertainty awareness. **Interpretability** makes clear what the system “knows” while **uncertainty awareness** reveals what the system does not “know.” This allows the user to **rapidly calibrate their trust** in the system's outputs, spotting flaws in its reasoning or seeing when it is unsure.”





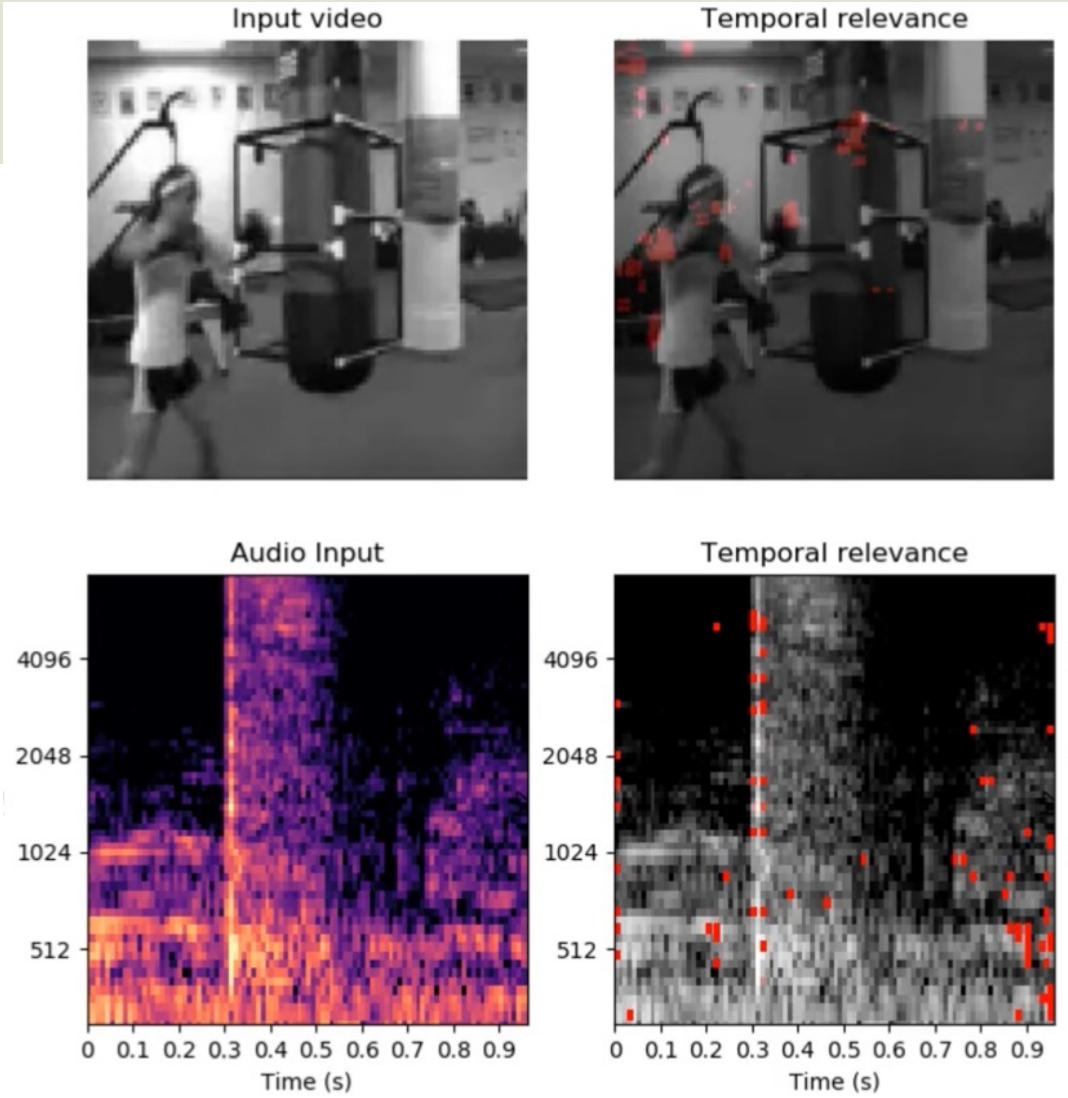
“Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI”
Patterns 2020 http://sl.dais-ita.org/science-library/paper/doc-5608

Coherent multimodal explanations for rapid trust calibration

We use **layerwise relevance propagation** to highlight most salient features in video and audio streams for a multimodal activity recognition classifier.

Intuition: “If it walks like a duck and quacks like a duck…”

CCTV operator concept: runs on edge devices in real-time, so explanations can be toggled on or off rather than asking “why?”



“VADR: Discriminative Multimodal Explanations for Situational Understanding”
Fusion 2020 http://sl.dais-ita.org/science-library/paper/doc-5538

ARL

[dstl]

Thanks for listening!

Credits: Katie Barrett-Powell, Dave Braines, Jack Furby, Mark Hall, Liam Hiley, Harri Taylor & Richard Tomsett

This research was sponsored by the U.S. Army Research Laboratory and the UK Ministry of Defence under Agreement Number W911NF–16–3–0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the UK Ministry of Defence or the UK Government. The U.S. and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon.

ARL

[dstl]