TAS-S
ANNUAL REPORT 2021-2022

UKRI AUTONOMOUS SYSTEMS NODE IN SECURITY

Foreword

The UKRI Trustworthy Autonomous Systems Node in Security (TAS-S) constitutes an exciting opportunity for collaborative research that is additionally supported by our unique security and Autonomous Systems (AS) test beds. The project has extensive stakeholder support, both domestic and international, from academics to AS providers, AS users and AS regulators.

Further details on all of the sections in this report can be found on the TAS-S website.

I am delighted to present the second TAS-S annual report.

These last twelve months have been a busy and exciting time for the TAS-S Node. We have established successful and productive partnerships across industry and academia, as well as growing our wide community of stakeholders. This report highlights just a few of our successes this last year, including details of our work with National Highways, our award for the Best Paper at MSN2022 in December, our published "Thought Piece" article produced with the TAS Hub and Thales, and our External Stakeholders Group Meeting 2022.

The Node's three research strands are producing some excellent results from their individual and collaborative research projects. Further details and posters outlining their recent work are included on pages 17-49.

I am very much looking forward to working with colleagues across the project and our wider network to continue to develop our activities further as we go forward into 2023.

Professor Neeraj Suri,
Principal Investigator, TAS-S
Lancaster University

Contents

Introduction

Autonomous Systems (AS) can be broadly categorised as the ability to effectively conduct a mission with varied levels of “absence of human intervention” including completely unsupervised operations. Typical examples, spanning an ever growing diversity of civilian, industrial and military applications across terrestrial, aerial and aquatic environments include autonomous vehicles, industrial automation, assisted living and a variety of logistical support to complement and supplement societal needs.

As technologically complex and networked cyber-physical entities, an AS needs to ensure “safe and secure” mission functionality despite the occurrence of any encountered cyber-physical disruptions. As such, an AS is a highly-dynamic entity that needs to adapt to the vagaries of its operational environments and security profiles (including changing threats). Providing “predictable, scalable and composable” security (of the AS assets, of the AS operations and the AS usage environment) in “uncontrolled and dynamic” operational environments is the objective of TAS-S.

The TAS Security Node’s research is centred around a seamless collaboration between fundamental cross-disciplinary security research and autonomous systems research at Lancaster and Cranfield Universities. To accomplish this vision, TAS-S utilizes interlinked cross disciplinary Research Strands (RS) to address 3 core challenge areas in autonomous system (AS) security:

Research Strand 1:
Securing the AS "usage" environment

Research Strand 2:
Can we secure the AS "operations" environment?

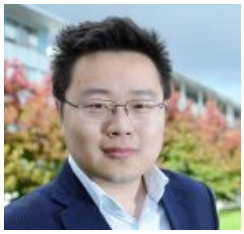
Research Strand 3:
Can we secure the AS "user" environment?

TAS-S Team

TAS-S assembles a cross-disciplinary team of internationally reputed security experts from Lancaster and Cranfield Universities who are based across a wide range of research areas including Distributed Systems, Controls, AI, Communications, Sociology and Law.

Research Strand Leads and Project Manager

Prof. Neeraj Suri,
PI, RS1 Lead
Lancaster University



Prof. Weisi Guo
Co-I, RS2 Lead
Cranfield University

Prof. Corinne May-Chahal
Co-I, RS3 Lead
Lancaster University

Pamela Forster
Project Manager
Lancaster University

Co-Is, Lancaster University

Prof. Plamen Angelov
Prof. Daniel Prince
Dr. Joe Deville
Prof. Catherine Easton

Co-Is, Cranfield University

Prof, Gokhan Inalhan
Prof. Antonios Tsourdos
Dr. Lisa Dorn

Postdoctoral Researchers, Lancaster University

Dr. Zhengxin Yu
Dr. Andrew Sogokon
Dr. Luke Moffat
Dr. Yi Li
Pierre Ciholas

Postdoctoral Researchers, Cranfield University

Dr. Burak Yuksek
Dr. Zhuangkun Wei
Dr. Oscar Gonzalez Villarreal
Dr. Anders åf Wahlberg

Affiliated PhD Candidates

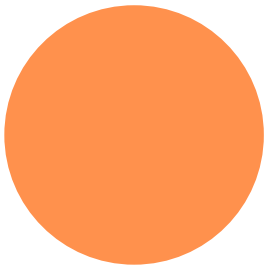
Xavier Hickman
Alvaro Lopez
Ovini Gunasekera
Julia Michelon Alvarenga

Governance

The node has an agile management structure to provide (a) efficient and responsive internal project management and (b) engagement with TAS Hub/nodes and external stakeholders. Further details can be found on the following pages or on the TAS-S website.

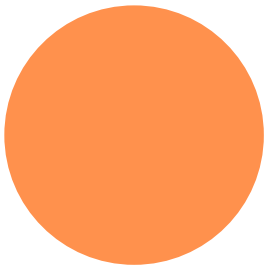
Project Management

The PI, Co-Is and Project Manager meet regularly as part of the Node's scheduled monthly "Research Activity Group (RAG)" and "Coordination Group (COG)" meetings. These groups have oversight of the day-to-day running of the project, plan engagement activities and events, and monitor progress with respect to the objectives, research outcomes and emerging risks.



Research Management

Each research strand has its own meeting structure, through which updates from each theme (2 or 3 per research strand) are discussed and opportunities for further collaborative work are identified. The Postdoctoral researchers from both Lancaster and Cranfield meet online every two weeks to present particular aspects of their research, suggest reading from the wider field, and discuss ways to work cross-research theme/cross-institution.



Governance (cont.)

Strategic Approach

The Advisory Group provides strategic advice and feedback on the project's research approaches, progress, quality and activity development. The Group has recently welcomed 2 additional members and now consists of 6 distinguished external stakeholders from academia and industry across the UK, Europe and the US. We are also in regular contact with the Hub Liaison Team to ensure that our Node works closely with the TAS Hub. Further details can be seen below:

Advisory Goup

Prof. Robin Bloomfield
Adelard

Prof. Phil Koopman
Carnegie Mellon University.

Dr. Hector Figueiredo
QinetiQ

Dr. Carl Sequeira
Flarebright

Dr. Arthur van der Wees
Arthur's Legal

Prof. Carl Landwehr
Center for Democracy & Technology, University of Michigan

Jo Marriott
EPSRC

Hub Liaison

Prof. Luca Vigano, King's College London
Prof. Derek McAuley, University of Nottingham
Prof. Jose Such, King's College London

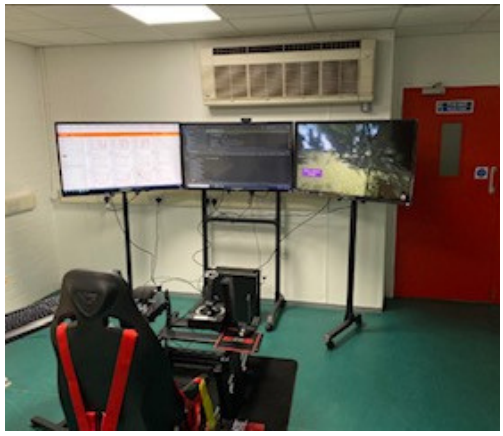
Testbeds

Lancaster and Cranfield Universities are home to specialist testbed facilities, including a unique autonomous systems test facility for combined air-ground vehicles at Cranfield.

Further details about our facilities can be found on the dedicated Testbeds page on the TAS-S website.

Simulation Environment for AI-aided Navigation of Autonomous Systems

The main aim of this simulation environment is to provide detailed mathematical model of the environment that autonomous systems are operating in . The system runs based on Unreal Engine and Airsim. It includes dynamical models of ground/aerial vehicles with sensor models such as inertial measurement unit (IMU), global positioning system (GPS), cameras and Light Detection and Ranging (LIDAR).



Real UAV Flight Arena

This is a real UAV experiment indoor platform called Arena in Building 83 of Cranfield University Main campus. The platform is assisted by Vicon system for UAV positioning and IMU and MARG data measurement. The aim is to run designed algorithms in real scenarios, and bridge the gap between academia and industry perspectives.

Testbeds (cont.)

Autonomous Systems Protocol Testbed

A testbed to replay, simulate, and benchmark autonomous systems communications supporting dynamic topologies, network state fluctuations, and highly scalable (2 to 10k instances).

Cyber Threat Laboratory

A partnership between Security Lancaster and Fujitsu Enterprise and Cyber Security, the lab is a collaborative platform that allows analysis of threats and behaviour to take place in a safe and controlled environment.

Lancashire Cyber Foundry

A series of multi-million pound secure digitalisation projects to help SMEs across Lancashire embrace innovations in digital and cyber technologies to defend, innovate and grow their business.

innovative Digital Infrastructure Defence (iDID)

iDID addresses pragmatic industrial requirements using applied research methods and focusses on cyber security and cyber threat intelligence for internet-enabled cyber physical systems.

Testbeds (cont.)

Multiuser Environment for Autonomous Vehicle Innovation (MUEAVI)

This outdoor test facility supports the rapid development of on and off highway. ground and airborne autonomous vehicles. The facility includes sensors with 4G/5G connectivity sensors, along with Lidar object recognition/test drone tracking on both LOS/NLOS basis.

National Digital Aviation Research and Technology Centre (DARTeC)

DARTeC is a £65 million facility integrating research and practice. It includes a fully functional airport, digital control tower, and air space control to offer a unique research and development environment.

National Beyond Visual Line of Sight Experimentation Corridor (NBEC)

NBEC provides a safe, managed environment to test and develop concepts , principles and related technologies to enable flying unmanned aircraft systems beyond visual line of sight in non-segregated airspace.

Website and Social Media

TAS-S website

We have further developed our dedicated website to detail all the different aspects of the Node including:

- **Node overview**
- **TAS-S Team**
- **Advisory Group/Hub Liaison**
- **Stakeholders**
- **Research Nodes**
- **Testbeds**
- **Publications**

Our website is a valuable resource not just for our Node but for the whole TAS network and the wider field. Therefore, it is regularly updated with details regarding the following:

- **Seminars** from TAS-S, Lancaster University, Cranfield University, and the other Nodes.
- **Events** from the TAS Hub and the wider field.
- **News** which is of interest to the Autonomous Systems Community.
- **Blogs** from our Project Manager and Researchers regarding the Node's latest activities and research.

A small sample of our current news posts and blogs can be seen on the right-hand side of the page. You can find out more by visiting our **website.**

Best Paper Award

Congratulations to colleagues Zhengxin Yu, Neeraj Suri, Plamen Angelov, and Yang Lu who were awarded best paper for 'PPFM: An Adaptive and Hierarchical Peer-to-Peer Federated Meta-Learning Framework' at MSN2022! A copy of the paper is

AI Bus carried first UK passengers (20th Jan 2023)

UK-based bus and coach operator Stagecoach has successfully transported its first passengers on board an autonomous bus during a trial in east Scotland.

Twitter and LinkedIn

Our social media accounts contain a huge range of information about upcoming seminars and events, latest job opportunities and news from across the TAS network and wider field **@TAS_Security**

Engagement: 2022

Despite the ongoing complications caused by the COVID-19 pandemic, the Node has organised and taken part in a range of online and hybrid events. These have included workshops and seminars, as well as engagement in the TAS Hub and other Nodes' initiatives. Further details can be found on the Node's website.

Event name/type	Description and impact
External Stakeholders' Group Workshop (ESG) 1st March, online.	• This full-day event provided the Node with a critical opportunity to network with around 40 TAS-S stakeholders, an international representation of prominent autonomous systems leaders from 30 academic, governmental and industrial organizations. • The third ESG workshop is scheduled for the 18th and 19th April 2023.
Researchers' Workshops 31st March-1st April 13th-14th July 20th & 21st September	• These events provided excellent opportunities for our postdoctoral researchers, PhDs and Co-Is from Lancaster and Cranfield Universities to meet each other face-to-face to discuss their research and explore collaboration opportunities. • These workshops have included researcher talks, poster presentations, demo discussions, training sessions on researcher communication and impact, and responsible research and innovation (RRI).
Engagement with the TAS Hub and other Nodes. Multiple dates.	• Workshop with the TAS Hub and Thales to develop a "thought piece" article based on Thales' use cases. This resulted in an article focussing on "Keeping Our Autonomous Systems Secure in an Uncertain World" and a short film highlighting the work of our node. • Wide-ranging involvement in the TAS All Hands Meeting (July 2022), including poster presentations and demo exhibitions.

Engagement: 2022 (cont.)

Event name/type	Description and impact
Engagement with the TAS Hub and other Nodes (cont.) Multiple dates.	• IEEE ACSOS Regional Event UK, presentations by the TAS Security and Verification Nodes (July 2022) • Prof. Neeraj Suri gave a presentation on TAS-S to the TAS Strategic Advisory Network. This network is made up of experts from policy, industry, academia and health (Feb 2022) • Invited talk to the International Geoprivacy Panel. • Prof. Catherine Easton presented a talk on "Cross-border Data Sharing in Emergencies: An Analysis of legal and ethical considerations" (Feb 2022).
Awards and Recognition (Multiple dates).	• Best paper award for Yu, Z., Suri, N., Angelov, P., and Lu, Y. (2022) 'PPFM: An Adaptive and Hierarchical Peer-to-Peer Federated Meta-Learning Framework' at MSN2022! A copy of the paper is available on our publications webpage. • Dr. Lisa Dorn has been one of the experts working on the Technical Committee (TC 241) to develop a new Standard on guidance on safety ethical considerations for autonomous vehicles from July 2019 to September 2022. This committee sits under the BSI Road Traffic Safety management systems of the International Organization for Standardization. This forthcoming ethical guidance standard is called ISO39003 and approval is expected by the end of 2023. • Invited talk to the Second IFIP Workshop on Intelligent Vehicle Dependability and Security (IVDS). Prof. N. Suri presented the talk "On Models and Reality" (July 2022).

Engagement: 2022 (cont.)

Event name/type	Description and impact
Stakeholder Seminar Series (#TASSTalks)	These online seminars were presented by our external stakeholders to capture their requirements, experiences and challenges. Around 25-30 external participants attended each of the following talks: • 01/03/2023: "*Trust and Governance for Autonomous Vehicle Deployment*", Dr. Phil Koopman, Carnegie Mellon University. • 18/03/2022: '*Trusted Data Sharing (TDS); sharing data based on trust in dynamic (eco)system life cycles*', Dr. Arthur van der Wees, Arthur's Legal. • 11/11/2022: '*Hierarchical Potential-based Reward Shaping from Specifications*', Dr. Dejan Ničković,, Austrian Institute of Technology. • 25/11/2022: ''An ML Boosted Software Engine for Next Generation Drones' Dr. Carl Sequeira, Flarebright. • 09/12/2022: ''*Discovering Unknowns on Visual Data*' Yang Zhou, Loughborough University.

Spotlight on: National Highways.

A key part of the TAS-S Node is the formation of a multi-disciplinary team to collaborate with stakeholders to explore the Ethical, Legal and Social Issues (ELSI) of Autonomous Systems (AS) Security (please see page 15) both in organisational and public contexts. The aim of such collaborations is to enable external stakeholders to reflect on the issues of ethics and security that relate their specific areas of activity. The ultimate aim of our work is to understand how organisations are navigating the challenges of designing and deploying AS in the UK, and to develop a series of broadly applicable resources and toolkits to support them in this regard.

In order to do this, our collaborative aims include the **identification of actionable insights** for stakeholder partners, that they can use in their strategies and/or their direct work with AS and the **creation of co-produced outputs**. Our role is not to act as ethical or security auditors, but to foster different forms of critical reflection, emerging out of the dialogue between practitioners and academics.

One of our Node's major collaborations over the past 12 months has been exploring how the ELSI approach can be applied to the work of National Highways both in organisational and public context with a focus on three case studies:

- **Connected and Autonomous Vehicles,**
- **Autonomous Plant and Construction,**
- **Supply Chain Challenges.**

As described more overleaf, we have utilised creative methods to map the issues of ethics and security relevant to National Highways which have included workshops, one-to-one interviews, and a forthcoming transitions report.

ELSI & National Highways.

ELSI

The Ethical, Legal and Social Issues (ELSI) framework is one of many cross-disciplinary approaches to technological innovation, which seeks to examine, address, and advise upon the wider implications of new technologies being implemented in society. In RS3, we draw on this framework to inform our research, our engagement with others within the TAS-S project, and our collaboration with external partners.

Ethical

While traditional ethical theories tend to focus on individual conduct and so here, on individual technological devices, the ELSI framework approaches ethics as an interconnected, complex process of negotiation, appraisal, and reflection. Who benefits from technologies being designed and used, and who is harmed as a result? Tools like Ethical Impact Assessment help developers in industry to audit their new technologies according to these values of benefit and harm.

Legal

Standardisation and best practice go hand in hand with a robust understanding of the legal landscape, as well as the capacity to change it. This requires opening channels between tech developers, operators, and policy advocates, so that legal practices can help forecast better AS futures, as well as responding to existing challenges.

Social

AS do not exist in a vacuum. They operate in, engage with, and respond to, pre-existing social structures and protocols. A key component of ELSI involves making space for communities to voice their thoughts, apprehensions, and desires for how AS work with and for them. Putting the ELSI framework into practice is not easy, nor should it be. But approaches like ELSI are essential for ensuring that new autonomous technologies are not only possible, but suitable for society. Future work in RS3 will also be using the Design Justice principles, in combination with ELSI, to look at ways of expanding the positives of AS according to community assets, and limiting harms imposed by techno-solutionism. You can find out more about Design Justice here

National Highways: Our Collaborative Approach

Stakeholder workshops.

The collaboration has included two workshops between TAS-S researchers and stakeholders from National Highways. The first workshop (January 2022) discussed ongoing projects National Highways have into connected autonomous vehicles (CAV), connected and autonomous plant, digital roads, public engagement, and cybersecurity. This led to a value mapping exercise to visualise the key priorities and issues National Highways are tackling, and how they connect. All parties shared an acute awareness of the need for proactive, ethically informed practices when it comes to designing, implementing and governing secure automation across National Highways' remit.

The second workshop (May 2022) allowed us to dive deeper into some key themes and areas of work which National Highways are currently managing; specifically around the two core themes of **Organisational Adaptation** and **Public Engagement** with reference to the three case studies outlined on the previous page, a particular focus on issues around security and ethics.

Feedback from National Highways:
'I leave our sessions wishing they were longer and looking forward to the next',
'[The workshop went] very well. Thought provoking - it's always good to run out of time, it means valuable conversations are taking place!'

Final Report

The final transition report is due to be delivered to National Highways in spring 2023. This will include actionable insights for them to apply going forward, as well as strategies for continuing to reflect in different ways about relevant issues of ethics and security. Further details about this very productive partnership can be found on the RS3 research webpages.

IPSOS Survey

RS3B and RS3C have also collaboratively designed a survey to explore how issues of AS security, in relation to CAVs in particular, are understood by the public. This resulted in over 400 responses with emergent insights indicating a high degree of concern amongst the public about CAV safety. Four online focus groups are now being carried out to add richer, qualitative data to the survey results, and examining with participants specific 'participatory backcasting' scenarios involving AS deployment on UK roads.

Research

The interconnectivity of our three research strands is illustrated in the image below. In addition, each strand is split into specialist themes focusing on a dedicated area of research. Further information about each of the research strands and themes can be found on the following pages and on the research pages of the TAS-S website.

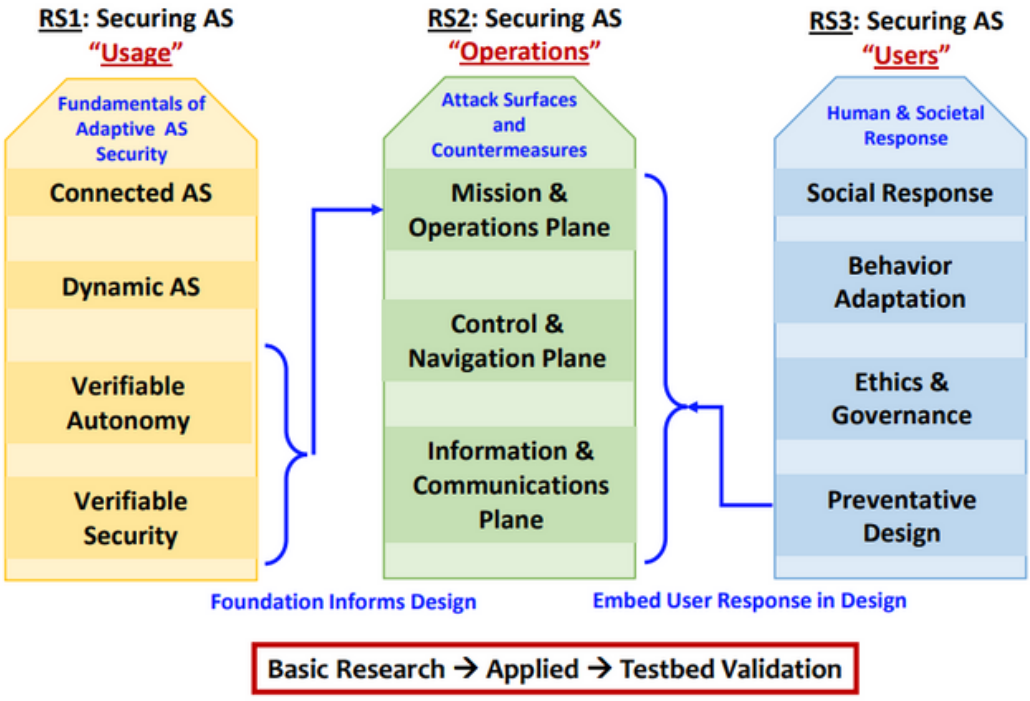


Fig. 1: The interconnectivity of the Research Nodes

RS1: Securing the AS "usage" environment.

To establish the fundamental AS "usage" framework for providing and assessing multi-layered, multi-dimensional adaptive AS security in dynamic mixed mode environments.

RS2: Can we secure the AS "Operations" environment?

To ascertain exposure (and their consequent mitigation) of AS "operations" to cyber-physical attacks by characterizing the attack surfaces (i.e. entry points and likelihoods) across the mission, control and information surfaces in a technology and mission-invariant manner.

RS3: Can we secure the AS "User" environment?

To ascertain the overall AS threats across multiple attacks, our approach tackles three interdependent AS surfaces (mission, control and communication), while the security foundations of RS1 and the human behaviour from RS3 are used to create holistic mitigation strategies.

TAS-S Overview: Bridging Gaps Between Makers and Users

Lancaster University, Cranfield University

Researchers: Dr. Anders af Wåhlberg, Pierre Ciholas, Dr. Oscar Gonzalez Villarreal, Dr. Yi Li , Alvaro Lopez, Dr. Luke Moffat, Dr. Andrew Sogokon, Dr. Zhuangkun Wei, Dr. Zhengxin Yu, Dr. Burak Yuksek.

UKRI TAS-S Trustworthy Autonomous System node in Security

Engineering and Physical Sciences Research Council

TAS-S

What is Autonomy?

The ability to effectively conduct a mission with varied levels of absence of human intervention including completely unsupervised operations.

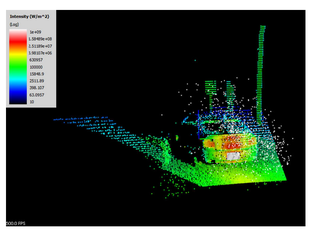
Autonomous System

- Perceive environment
- Make decision
- Actuate a movement

Coordination
- Homogenous / heterogenous fleet operations.
- Resource sharing between assets.
- Maximising the operation effectiveness, safety and security.

Learning Enabled
- Acquiring environmental data and behavioural adaptation in real-time.
- Provable safety and robustness.

Dynamic
- Real-time adaptation capability.
- Responsive for environmental, dynamical and operational variations

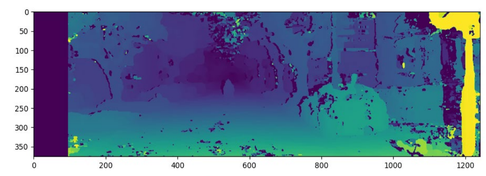


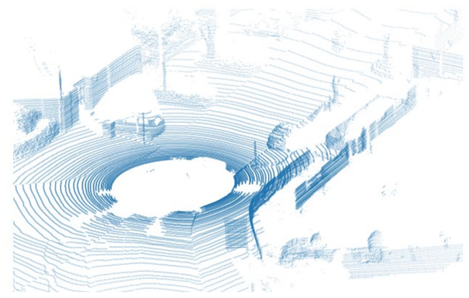
What is trust is performative?
From lack of alternatives
What ways can trust be done differently?

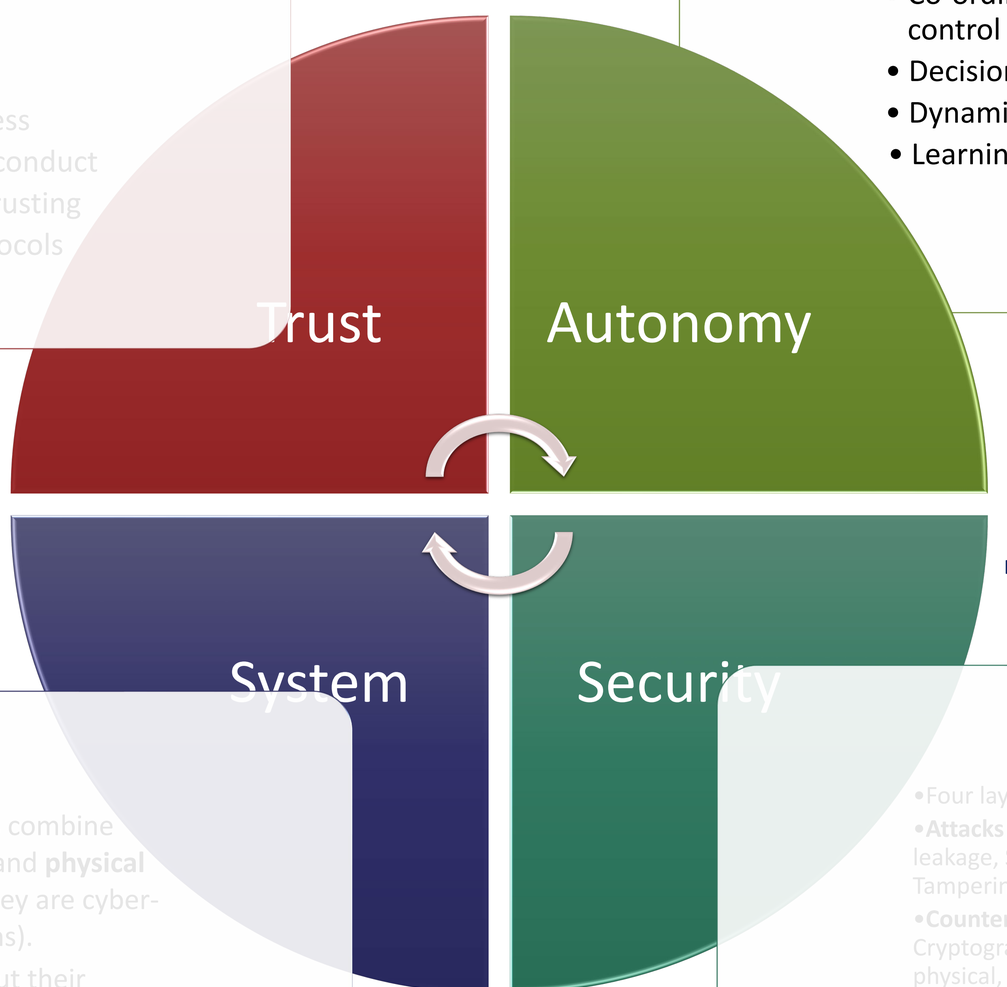
How and/or why do publics trust processes such as this?

Factoring in resistance and dissent

- Expectation meeting reality
- Risk perception
- Verifiability
- Societal Readiness
- Trust as ethical conduct
- Other ways of trusting
- Indigenous Protocols

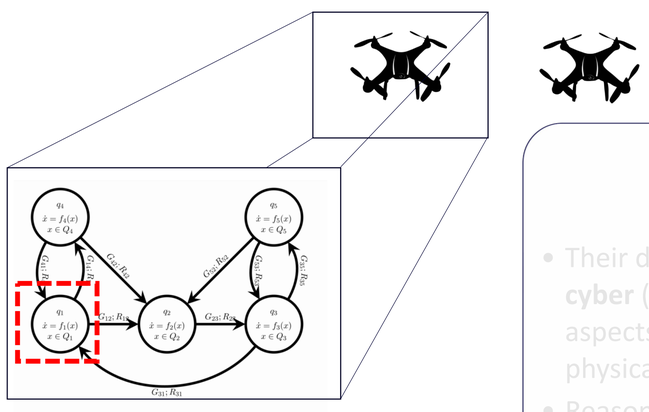


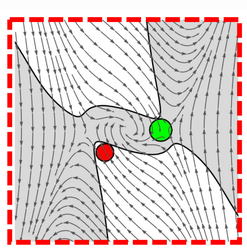




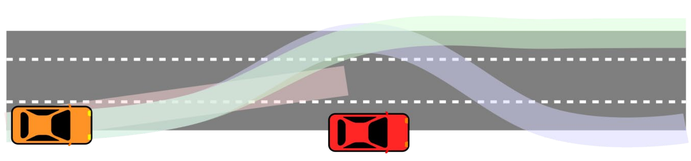
- Co-ordination & control
- Decision making
- Dynamic
- Learning enabled

- Autonomous systems are often **networked** and operating in environments where they are exposed to attacks.

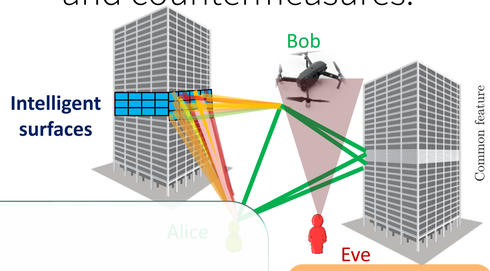


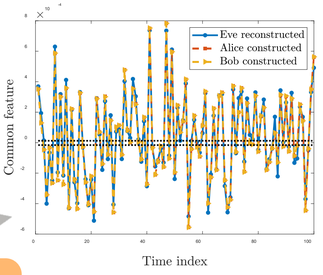


- Their dynamics combine **cyber** (digital) and **physical** aspects (i.e. they are cyber-physical systems).
- Reasoning about their physical behaviour is a big challenge.



- Security of Autonomous systems are from different layers, with specific attack vectors and countermeasures.





- Four layers of security
- **Attacks vectors** – Information leakage, Spoofing, Jamming, Tampering of privilege
- **Countermeasures**- Cryptography, (quantum, physical, mathematical), authentication, Federate differential privacy

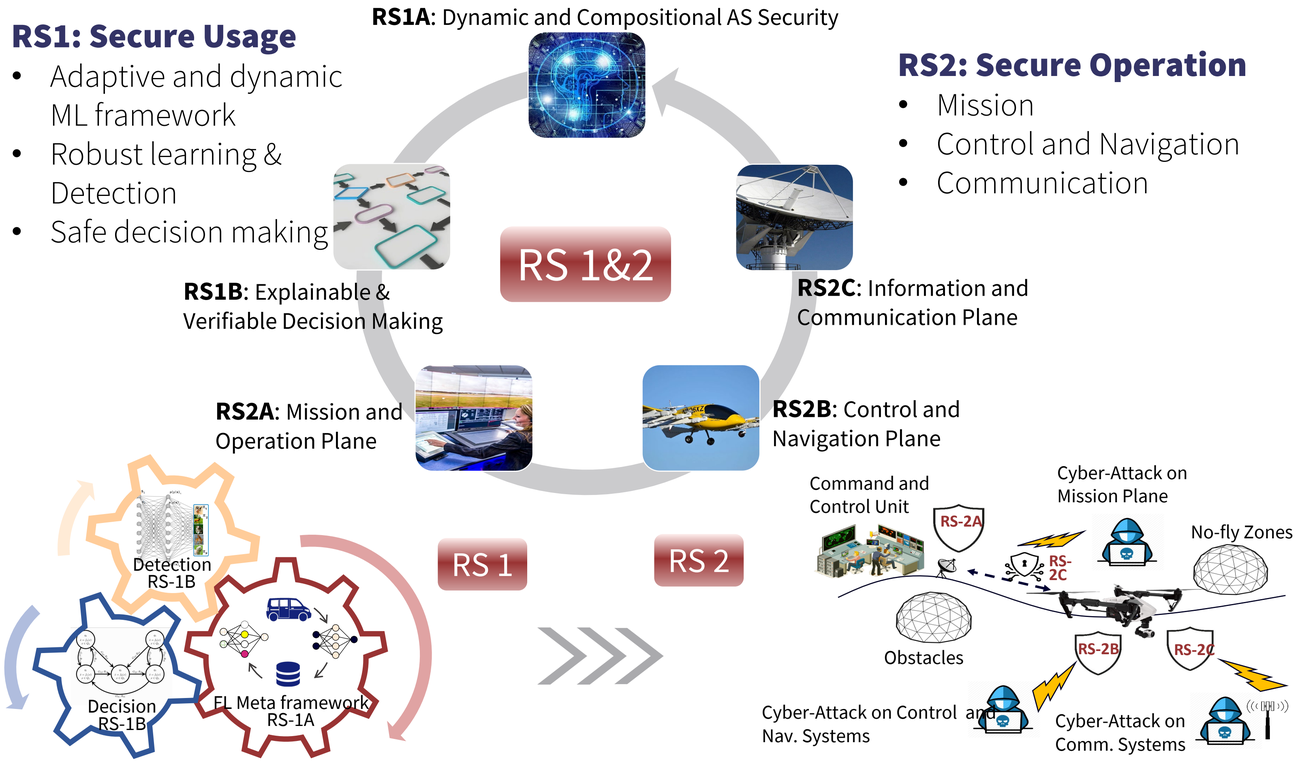
Timestamp	User ID	Entropy	Access	MAC
Clear		Encrypted (KD/OTP)		Clear

App Layer	Learning based method to prevent malicious uses, protect data privacy
Network Layer TCP, UDP	Internet protocol and socket security
Access Layer Virtual, Mac	Authentication against spoofing and sniffing
Physical Layer: control, comms	Protect channels from contamination (jamming, spoofing), and prevent information leakage

Secure Usage and Operation of AS (RS1 & RS2)

Secure Operation and User of AS (RS2 & RS3)

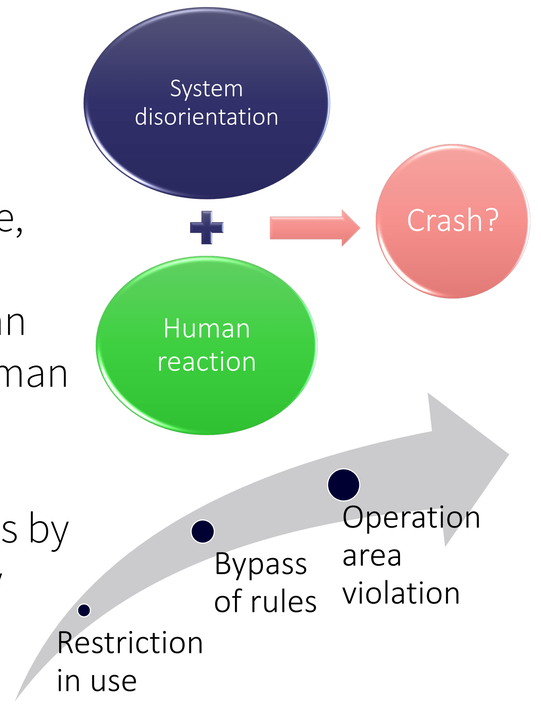
RS1: Secure Usage
- Adaptive and dynamic ML framework
- Robust learning & Detection
- Safe decision making



RS2: Secure Operation
- Mission
- Control and Navigation
- Communication

RS3: Secure User
- Individual Behavioural Adaptation
- Organizational Processes
- Ethical & Legal Security Ecosystem

- If an attack disorientates a vehicle, how does the human react?
- How should the system react to an attack situation to enable the human to safely handle the situation?
- Can autonomous systems be misused through violation of rules by provision of faulty information by users?



Lancaster University
Cranfield University
UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Research Activities: RS1A

RS1-Theme A: Dynamic and Compositional AS Security

Lead: N. Suri. Participants: A. Tsourdos, G. Inalhan, A. Sogokon

Overview

The research being carried out by RS1A addresses the fundamental challenges of specifying AS interfaces and the emergent security properties over compositions across AS and/or with the environment and especially adaptivity that characterizes AS operations.

Our intent is to develop a conceptual framework characterising the relationships across security attributes and the role of collaborative, disruptive and scalable security composition in AS, along with a run-time security policy framework for AS.

Research activities

RS1A activities have progressed across two dimensions, namely (a) formal specification of AS operation/safety and (b) development of robust ML techniques to support AS functionality.

Specifications for Autonomous Systems: Specifying the intended behaviour of AS is essential to establish a reference baseline to ascertain the type and degree of any violations that may compromise safety and security. The work compiles potential approaches and highlights open issues in formal specifications for AS's.

Improving Tool Support for Continuous Systems in the TLA+ Toolbox: The established TLA+ framework lacks support for checking invariance properties essentially needed for continuous systems such as ASs. We are developing an open-source implementation of an invariant checker that can support TLA+ as an external to provide automation in safety proofs.
Robust Controlled Invariance of Sets (Methods for Checking and Generation)In order to design a system that is resilient to faulty input attacks, we develop techniques to guarantee that the behaviour of the system satisfies its safety specifications.

PPFM: An Adaptive and Hierarchical Peer-to-Peer Federated Meta-Learning Framework: We have developed a dynamic ML approach, where a distributed and defragmented federated meta-learning architecture adaptively tunes itself to match varying data characteristics by utilizing multiple learning loops in dynamic distributed AS environments.

Research Activities: RS1A

Research activities cont.

RAFL: A Robust and Adaptive Federated Meta-Learning Framework against Adversaries. RAFL reduces the impact of adversarial model updates in AS environments. RAFL leverages rule-based detection approaches to identify and remove adversaries.

PINCH: An Adversarial Extraction Attack Framework for Deep Learning Models. This empirically driven research focuses on discovering the relationship between adversarial Machine Learning attack effectiveness and Deep Learning model characteristics.

Looking Ahead...

RS1A has the following research activities planned for the next six to twelve months:

- Extension of AS specifications to cover communication (RS2) and Human Factors (RS3)
- Collaboration with RS2 to utilise PPFM/RAFL for this problem to navigation control.
- Development of TLA+ support toolbox

Highlights & selected publications

- Input on AS datasets and specifications with Airbus, TTTech, and BAE Systems.
- PC member; ACM Intl Conference on Hybrid Systems: Computation & Control, 2022 – ACM HSCC ’23 (Andrew Sogokon)
- ECRs Co-Chair; TAS Symposium 2023 (Zhengxin Yu)
- Publicity Chair; IEEE Intl Conf on Metaverse, Computing, Networking & Apps – IEEE Meta-Com 2023 (Zhengxin Yu)
- Publicity Chair; IEEE Intl Conf on Cloud Networking - IEEE CloudNet-2022 (Zhengxin Yu)
- PC Member; ACM Middleware 2023 (Neeraj Suri)
- PC Member, IEEE DSN 2023 (Neeraj Suri)
- Associate Editor, IEEE Trans on Big Data (Neeraj Suri)
- Associate Editor, ACM Computing Surveys (Neeraj Suri)
- Yu, Z., Lu, Y., Angelov, P. & Suri, N. 'PPFM: An Adaptive and Hierarchical Peer-to-Peer Federated Meta-Learning Framework', IEEE Intl conference on Mobility, Sensing and Networking (IEEE MSN-2022), 16th December 2022.
- Yu, Z., Yuksek, B., Suri, N & Inalhan, G. Federated Meta Reinforcement Learning for UAV Navigation in Urban Airspace, poster at the Safe and trustworthy AI workshop, TAS 2022
- Sogokon, A., Yuksek, B., Inalhan, G. & Suri, N (2022), Specifications for Autonomous Systems, ArxivX,

Research Activities: RS1B

RS1-Theme B: Explainable and Verifiable Decision Making for AS Security

Lead: P. Angelov. Participants: N. Suri, W. Guo. G. Inalhan, Z. Yu, A. Sogokon, Y. Li, O. Gunasekera.

Overview

RS1B aims to address two research challenges. Firstly, the control behaviour of an AS is often non-deterministic as an AS adapts to changes in the operational environment, resources, sensory streams and objectives to yield an "optimal" solution. This nondeterminism makes verification of the security attributes unviable by classical testing and verification approaches that, typically, verify a given static property. This is a standalone challenge as autonomy, usually, results in non-deterministic outcomes unable to support reproducibility of scenarios and results. Secondly, AS operate on data streams from sensory inputs for analysing data related to the mission, situation awareness, the navigation, and control. This results in the use of data-driven reasoning techniques.

Our intent is to develop dynamic verification methodologies, explainable-by-design DL architectures that lend themselves to reasoning interpretation as well as to visualization, and symbolic surrogate models for DL-based automation reasoning techniques.

Research activities

With regards to technical progress, RS1B has proposed several AI-driven methods to solve the research challenges in dynamic verification and validation.

Firstly, we proposed an adaptive machine learning framework to improve security and achieve resilience in autonomous systems. The framework defragmented machine learning models where hierarchical loops can provide dynamic interaction when attacks on machine learning occur.

Secondly, we proposed a robust-by-design algorithm, Sim-DNN, which detects adversarial attacks through the similarity-based mechanism. Different from the conventional defense mechanisms such as BaRT and ResUpNet, the Sim-DNN demonstrated outstanding performance even when suffering from adversarial training.

With regards to conceptual progress, we performed a thorough analysis of existing adversarial attacks, defence mechanisms and attack surfaces as well as the possible implications within autonomous systems aiming to validate the techniques being developed.

Research Activities: RS1B

Looking Ahead...

RS1B has the following research activities planned for the next six to twelve months:

- Attacks from real-world scenarios are often unknown during the training stage. Although, some of them can be expected, many more attacks come from unexpected domains such as the unpredictable environment or unexpected intruders/objects. Therefore, the generalisation ability of any well-trained model is critically important. A paper entitled, “Few-Shot Imperceptible Adversarial Attack Based on Domain Adaptation” is under development to address cross-domain adversarial attack detection in autonomous systems using machine learning tools. Different from conventional domain adaptation methods, we plan to decompose a network into two components, i.e., feature extractor and detector. In the first stage, we train one feature extractor with different detectors and datasets to boost the generalization ability of the feature extractor. In the second stage, we train one detector with different feature extractors and datasets. Furthermore, in the testing stage, the well-trained feature extractor and detector are combined within the final model to detect the adversarial attack and mask the attacked pixels.

Highlights & selected publications

- E. Soares, P. Angelov, and N. Suri, ‘Similarity-based deep neural network to detect imperceptible adversarial attacks’, IEEE Symposium Series on Computational Intelligence, IEEE SSCI, IEEE Xplore, 2022 and also presented at 2022 IEEE Symposium Series on Computational Intelligence SSCI 2022, 4th – 7th December 2022, Singapore
- Work with QinetiQ, 2ExcellGeo Ltd
- IEEE Standard on explainable AI, P2976, initiator, sponsor and WG lead initially and now a member of the WG (Prof. P. Angelov).

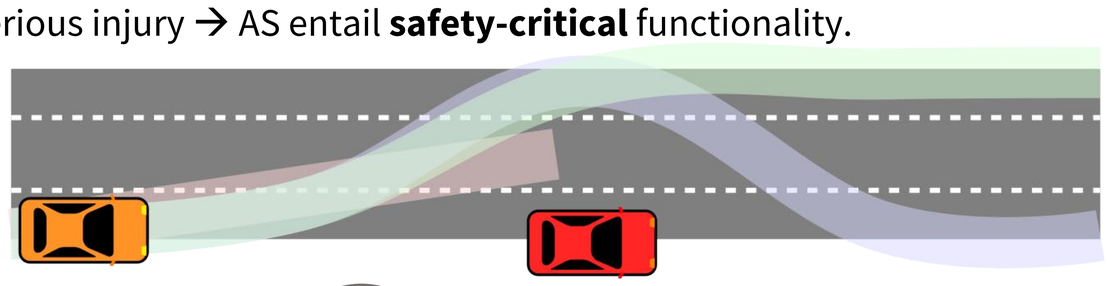
RS1: Securing the Autonomous System Usage Environment

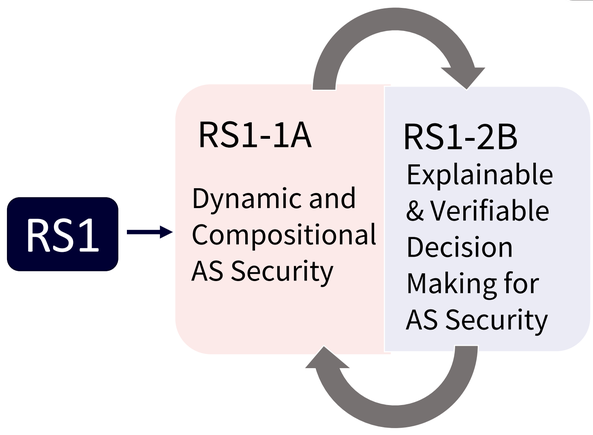
Lancaster University

Researchers: Dr. Andrew Sogokon, Dr. Zhengxin Yu, Dr. Yi Li
Investigators: Prof. Neeraj Suri, Prof. Plamen Angelov

RS1 - Secure Usage of Autonomous Systems

Autonomous Systems (AS) are typically *Cyber-Physical Systems* (CPS) where malfunctions can lead to catastrophic consequences, such as loss of life or serious injury → AS entail **safety-critical** functionality.





Challenges in RS1

- AS operates in dynamic, unpredictable environments.
- AS may be exposed to attacks; attacks are often complex – may be discrete, collusion and multi-layered.
- AS often process large amounts of data with complex data structures.
- AS needs to make "adaptive & run-time" decisions with *incomplete and uncertain* data streams & resources.
- AS nodes are *mobile*.

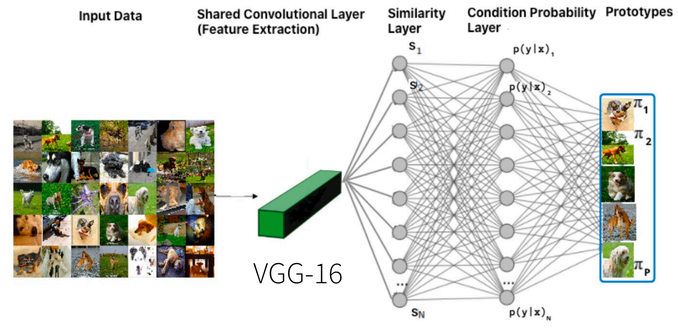
Engineering and Physical Sciences Research Council

TAS-S

RS-1B (2): Detecting Imperceptible Attacks

- Similarity-based Deep Neural Networks (**Sim-DNN**) can be used to detect *imperceptible adversarial attacks* on the sensors (e.g. vision system) of AS.

Sim-DNN



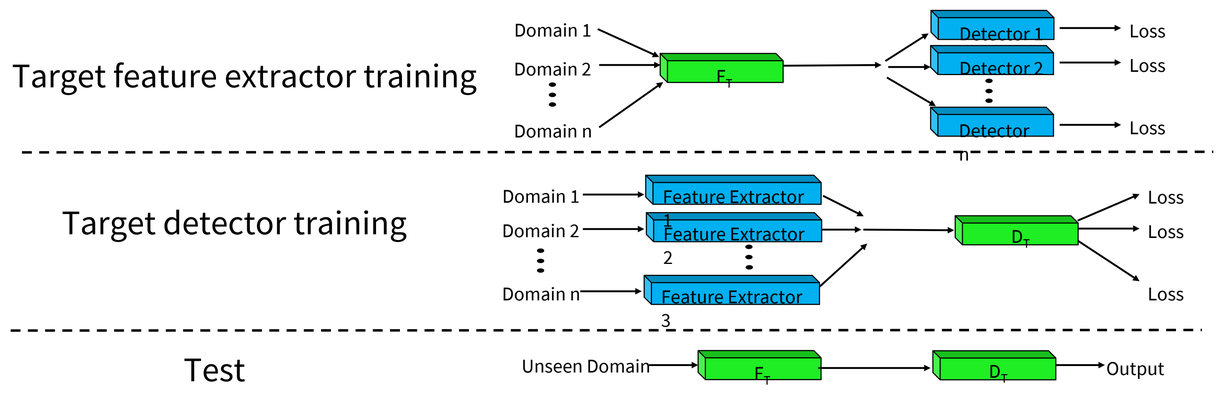
Pros:

- These frameworks provide excellent results for various attacks.
- These methods require few manual-engineering.

Cons:

- Weak adaptability and transferability to new domains, e.g., attacks or datasets.
- Slow training due to large model scales, particularly for the feature extractor (VGG-16).

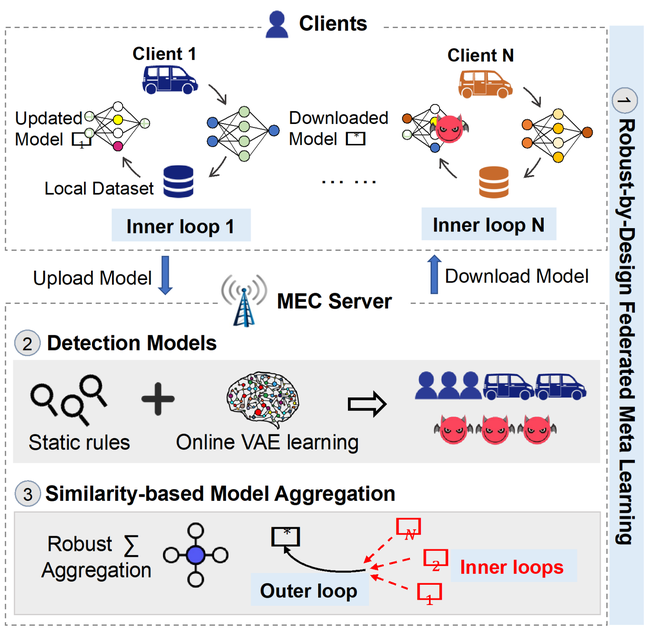
RS-1B (2): ML Domain Generalization Framework



- The feature extractor or detector is trained with a partner who is well tuned for different domains.
- In the test stage, the trained target feature extractor and detector are combined with the FFN to detect attacks in unseen domains.

RS-1A: RAFL- Dynamic & Compositional AS Security

- Develop a **robust and adaptive** federated meta-learning framework (**RAFL**) resilient against adversaries.



Goals:

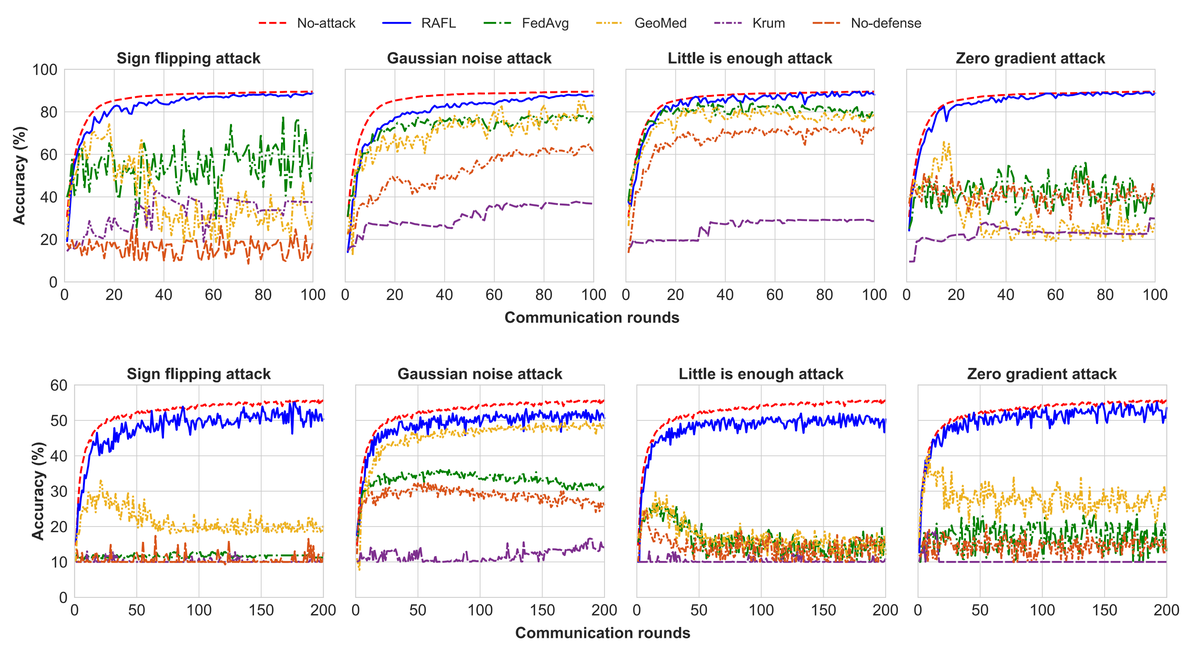
- Leverage distributed AS nodes to collaboratively train a global model to quickly adapt to new environments.
- Defend against adversarial attacks to reduce negative impact of attacks on ML models.

Key techniques:

- Federated meta-learning: Decentralized inner/out loops to train ML models.
- Rule-based and Variational Autoencoder (VAE) online learning-based detection model to detect adversarial attacks.
- A similarity-based model aggregation to conduct a global meta-model to further reduce the likelihood of uploading adversarial models from AS nodes.

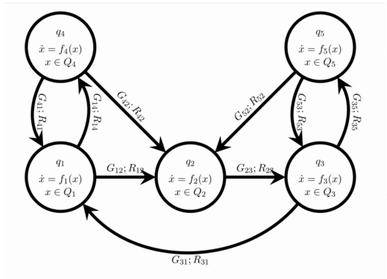
RS-1A: RAFL- Experimental Results

RS-1A: The experimental results demonstrate that the proposed **RAFL** framework is robust by design and outperforms other baseline defensive methods against adversaries in terms of model accuracy and efficiency.



RS-1B (1): Safe Decision Making in AS

- Establishing **safe and secure** operation of an AS in uncertain and dynamic environments is part of the focus of our research in **RS-1B** (*Explainable and Verifiable Decision Making*). We have undertaken a survey of specifications of AS, focusing on *formal specification*.
- Formal modelling and verification of CPS is highly challenging, but can help in providing very strong guarantees about the behavior of AS.



- We are working towards adding support for reasoning about CPS in the formal verification framework of **TLA+** based on Lamport's Temporal Logic of Actions.
- Formal methods can provide *verifiable solutions* to trustworthy decision making in AS.

RS-1B (1): Safe Decision Making in AS

RS-1B(1): We have implemented a *proof obligation generator* for checking continuous inductive invariants (the proof obligations are discharged using the SMT solver **Z3**) and are currently engaged in integrating it with the **TLA+ Toolbox**. Enables a convenient way of proving safety of continuous systems within the formal framework of the TLA+ Toolbox and will support formal verification of CPS.

RS-1A & 1B: Ongoing Work

- RS-1A: Develop a mobility-aware adaptive machine learning framework
- RS-1B (1): Formal specification of AS Safety and Security
- RS-1B (1): Case studies of safety verification of CPS in the TLA+ Toolbox. Integrate proof obligation generator into the Proof Manager in TLA+ Toolbox.
- RS-1B(2): Visualization results of the proposed algorithm will be completed.
- RS-1B(2): Adaptability and transferability will be evaluated in real-world photos.

Lancaster University

Cranfield University

UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Enabling Formal Safety Verification of Cyber-Physical Systems in TLA+

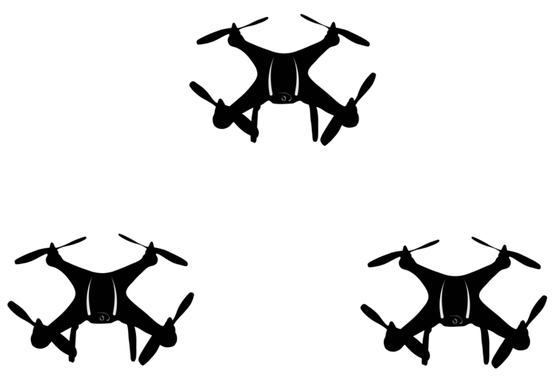
Lancaster University

Researcher: Dr. Andrew Sogokon
Investigator: Prof. Neeraj Suri

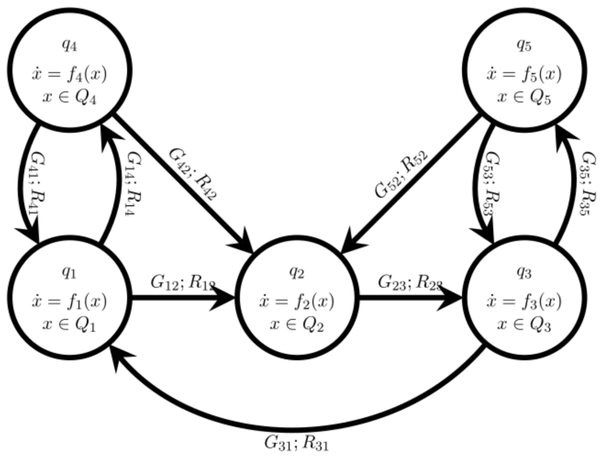
Safety-Critical Cyber-Physical Systems

Cyber-Physical Systems

- Cyber-Physical Systems (**CPS**) combine discrete and continuous behaviour.
- Examples include digital computer systems that operate in a continuous physical environment.
- Some CPS are **safety-critical** which means that failures can result in catastrophic consequences.
- Examples of safety requirements for CPSs include *collision avoidance* between autonomous vehicles in the aerial as well as the terrestrial domain.



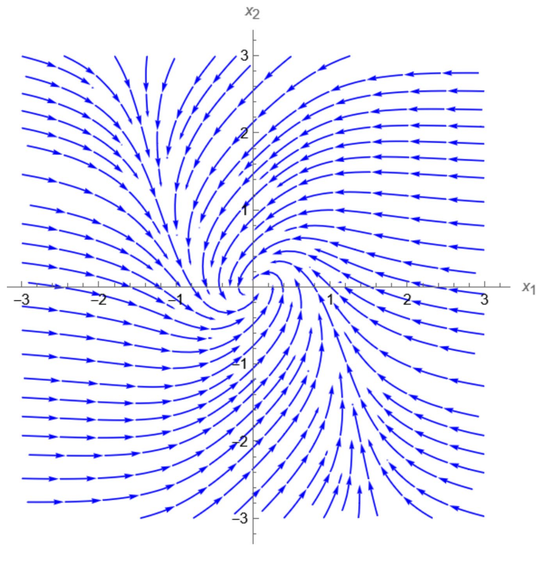
Formal Models of CPS



- Cyber-Physical Systems can be represented formally, e.g. using operational models such as hybrid automata or hybrid programs.
- A formal model of a CPS provides a mathematically precise description of the system that can be rigorously analysed.
- For **safety-critical** CPS it is important to ensure that the system adheres to its safety specification (e.g. avoids collisions at all times).
- A formal model of a CPS can (in some cases) be checked against a formal safety specification (typically stated using a formal logic). If successful, the safety of the model can be rigorously established.

Continuous Dynamics of CPS

- Continuous behaviour in CPS is usually governed by systems of ordinary differential equations (ODEs).
- Geometrically, a system of ODEs corresponds to a vector field defined on n-dimensional Euclidean space (where n is the dimension of the system).
- Solving ODEs is usually not possible analytically.
- Non-linear ODEs are particularly difficult to analyse.



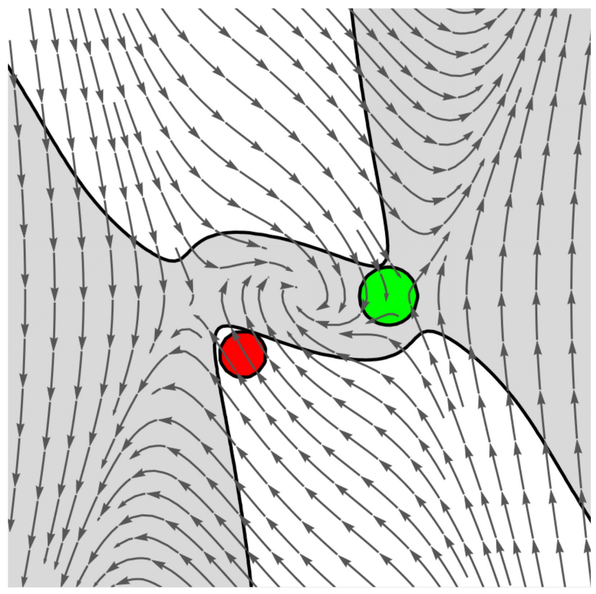
Engineering and Physical Sciences Research Council

TAS-S

Safety Specifications for Continuous Systems

Safety Specifications

- A **safety specification** for a given system requires two elements:
  - 1 - A description of the possible initial states from which the system may begin its operation.
  - 2 – A description of undesirable (i.e. unsafe) states into which the system must never transition.

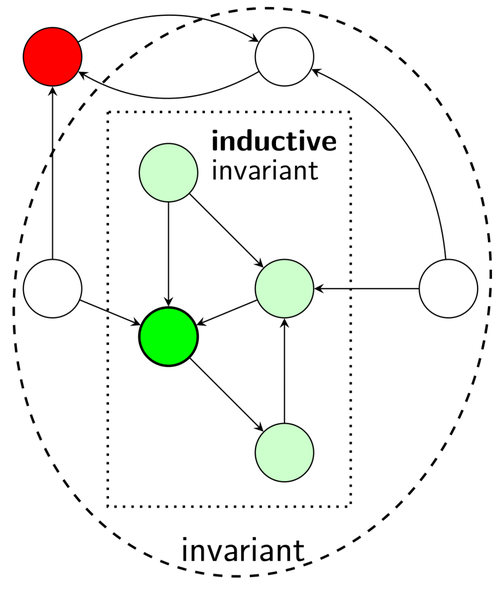


- **Safety verification** is concerned with proving a safety specification, i.e. rigorously demonstrating that a system may never transition into any of the unsafe states provided that it starts operating from one of the specified initial states.

Formal Verification in TLA+

Temporal Logic of Actions

- Lamport's Temporal Logic of Actions was designed to enable formal modelling and verification of concurrent systems. It enjoys excellent tool support in the form of the **TLA+ Toolbox** and has been successfully applied in industry.
- Formally proving safety specifications of discrete transition systems is typically done by finding an appropriate **invariant**.



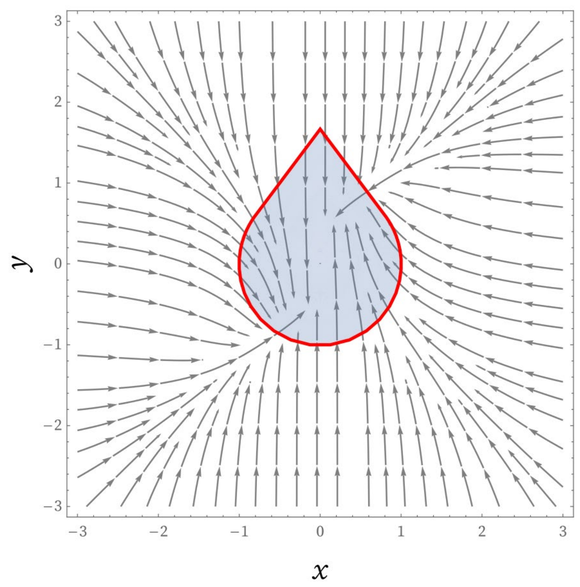
Inductive Invariants

- An **invariant** is a set of states that:
  - It includes all the initial states (as described in the safety specification).
  - It does not include any of the unsafe states.
  - The unsafe states are not reachable from the initial states.

An invariant is **inductive** if there are no transitions out of the invariant.

Checking Continuous Inductive Invariants

- A corresponding notion to an inductive invariant in continuous systems is that of a **positively invariant set**.
- There is a rich theory and powerful results about positively invariant sets in dynamical systems.
- More recent work in computer science has established that it is **decidable** to check whether a set is positively invariant (provided it is described using polynomial functions).
- This result makes it possible to perform safety verification without having to solve the ODEs.
- Adding support for checking continuous invariants would greatly facilitate CPS verification in the TLA formal framework.



Lancaster University

Cranfield University

UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

RAFL: Robust Federated Meta Learning Framework Against Adversaries

Lancaster University

Researcher: Dr. Zhengxin Yu
Investigators: Dr. Yang Lu, Prof. Neeraj Suri

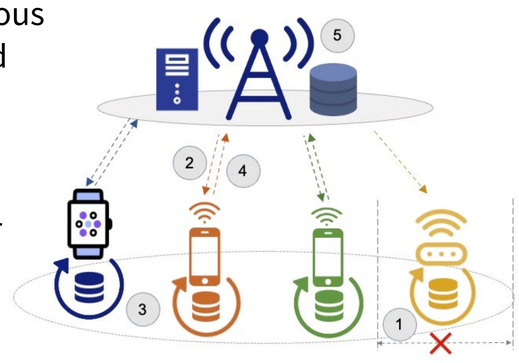
Federated Learning (FL)

FL is capable of leveraging distributed personalized datasets from multiple clients to train a shared global model in a privacy-preserving manner

Problem: FL systems can be vulnerable to various kinds of failures and attacks (data poisoning and model poisoning).

➡ **Degrade the learning performance of FL**

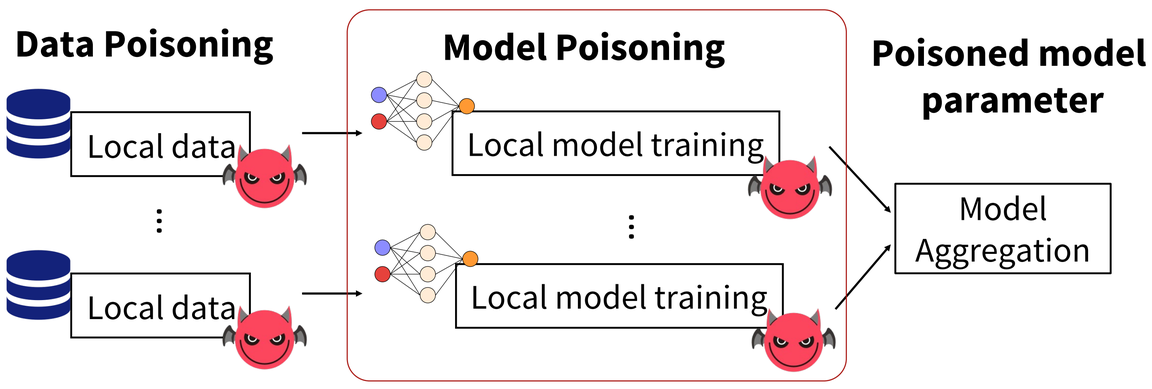
Impact: reduce model accuracy, quality of user experience, trustworthiness, resilience and communication overhead



SOTA: Robust learning and adversarial client detection

Challenges:

- Clients upload unreliable model updates intentionally or unintentionally.
- Local resource heterogeneity (Non-IID data distribution)
- Attacks are complex –discrete, colluding, multi-layered, moving-target behavior
- Dynamic environments (mobility, join-leave behavior, etc.)

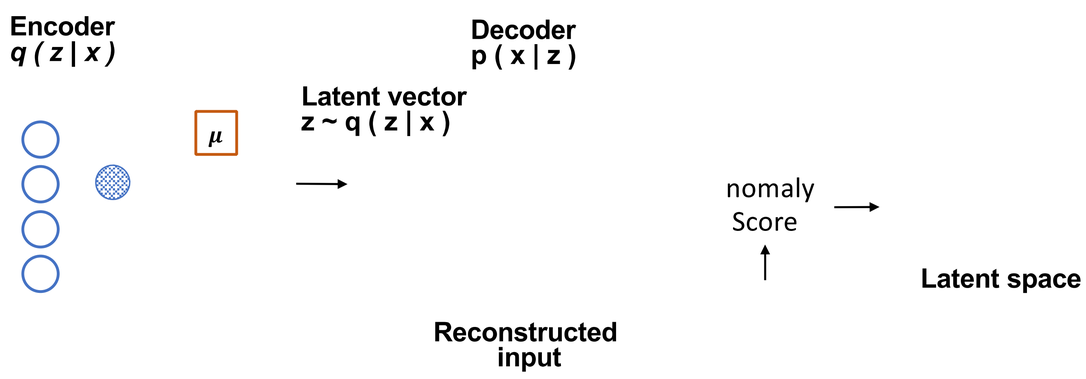


Robust Federated Meta Learning Framework

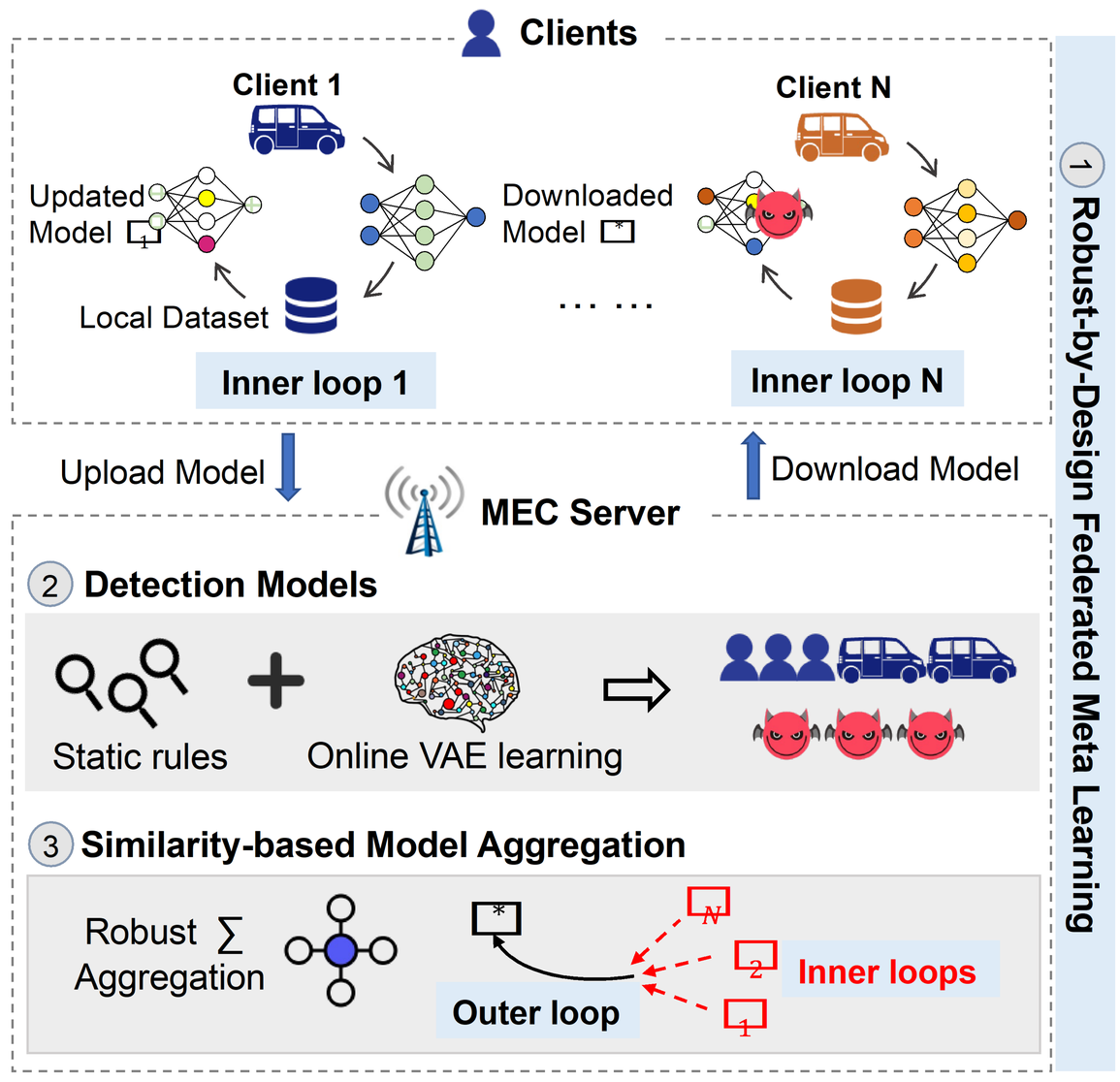
Develop a **robust and adaptive** federated meta-learning framework (RAFL) against adversaries

Contributions :

- A robust-by-design federated meta-learning architecture is proposed to adaptively defend against a range of adversarial attacks.
- A composite rule-based and learning-based detection method is developed to effectively identify adversarial clients via ranking domain and low-dimensional embeddings.
- An adaptive model aggregation method is proposed to aggregate the global model by considering the degree of similarity between the meta-model and calculated mean model to resilience attacks.

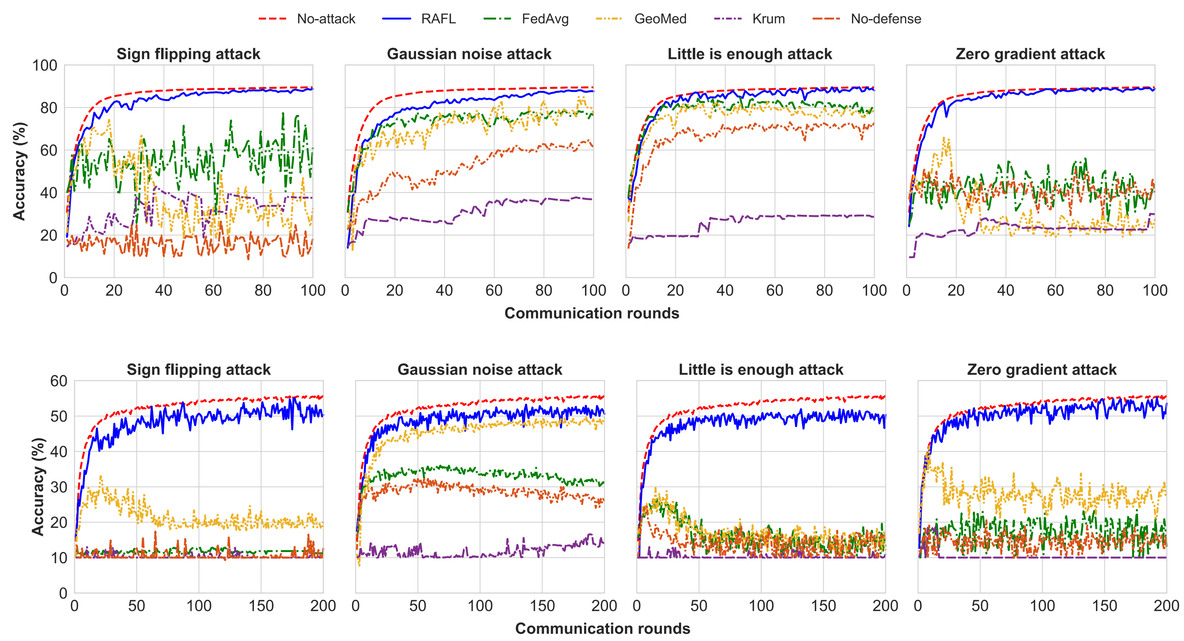


RAFL System Model

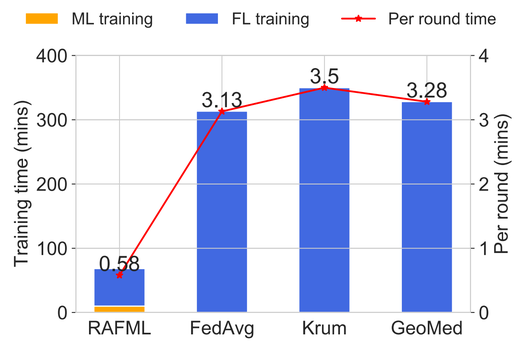


Experimental Results

The experimental results demonstrate that our proposed RAFL framework is robust by design and outperforms other baseline defensive methods against adversaries in terms of model accuracy and efficiency.



We compare RAFL's training time with other benchmark defence schemes. Total training time of RAFL (detector, FL training time) is less than SOTA.



Conclusion

- We have proposed a robust FL framework against adversaries, which combined a rule-based detection method and an online learning-based detection method to effectively distinguish adversarial clients from benign clients.

Future Work

- Explore the applicability of the RAFL to multi-attacks and consider more advanced ML models
- Develop a mobility-aware adaptive federated meta learning framework

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

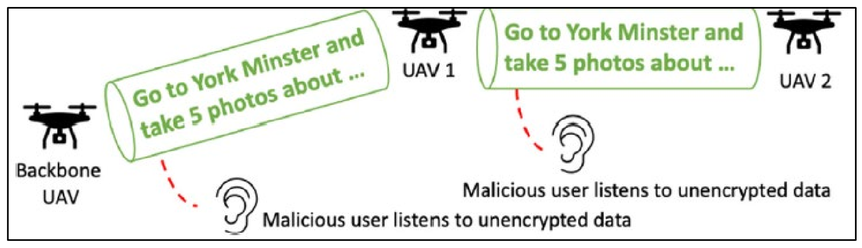
Parameterised Verification of Security Properties in Distributed Cyber-Physical Systems

Lancaster University

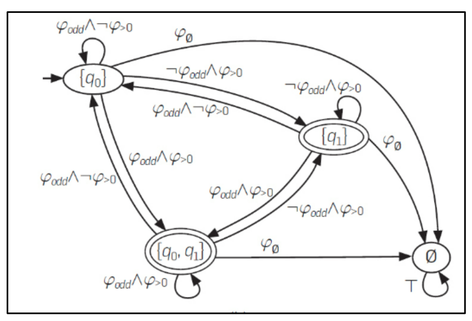
Researcher: Ovini Gunasekera
Supervisors: Dr Antonios Gouglidis, Prof. Neeraj Suri

Motivation: Security Attacks in Unmanned Aerial Vehicles

- Unmanned-Aerial Vehicles (UAVs) are used for persistent intelligence, surveillance, reconnaissance, targeting military personnel
- Adversaries attempt to interfere with the *Confidentiality, Integrity, Availability (CIA)* of drones affecting their overall functionality to either manipulate stored data or crash one or more drones of the swarm
- UAVs dynamically join and leave a swarm, therefore need to ensure the security of a swarm of an **arbitrary** number of drones at all times
- Swarms are increasingly developed to accommodate more UAVs for more complicated tasks, thus need for a **scalable** technique to detect/mitigate cyber-attacks on swarms of UAVs
- Example of security attacks
  - Battery Depletion and Exhaust Energy Attacks
  - Eavesdropping Attacks
  - Spoofing and Jamming Attacks



Parameterised Verification



- **Parameterised verification** problem is to verify (prove) some property regardless of the number of participants involved
- Towards solving the **state explosion problem** of automated verification i.e. model checking
- Non-parametric vs. Parametric verification: Parametric verification verifies all instances of the system at once
- Fully automated techniques for parameterised verification
  - Infinite auxiliary constructs: Automates user-supplied interactive steps in deductive verification
  - Counter abstraction: Abstraction of parameterised systems into a finite-state system and abstraction of the property to be verified
- Parameterised verification of multi-agent systems
  - Cut-off identification procedure: Number of components sufficient to analyse when evaluating a specification

Towards Parameterised Verification of Distributed CPS

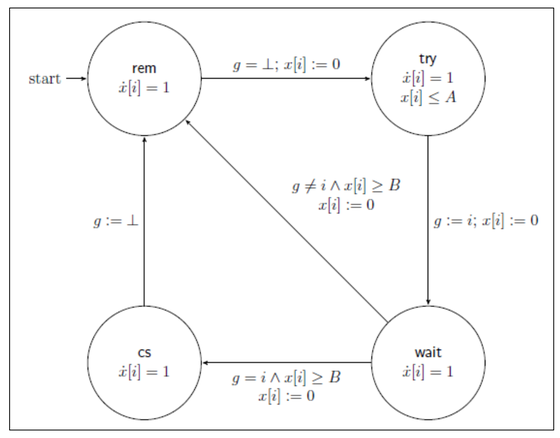
- Distributed CPSs are naturally parameterised by the number of participants (CPSs) involved
- Enables **scalable verification** of CPSs which are increasingly built and used
- Reasoning about **unbounded systems** where the number of components are not known during design time
- Need to automatically verify that any interaction between any arbitrary number of CPSs does not violate a property

Engineering and Physical Sciences Research Council

TAS-S

Specification of Cyber-Physical Systems

- **Hybrid automaton** is a formal model for a mixed discrete-continuous system
- Used in safety-critical applications i.e. CPSs

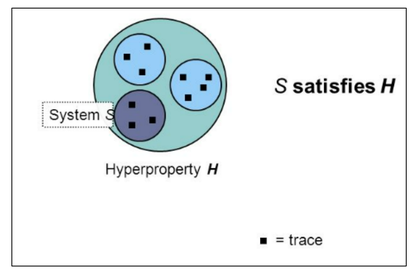


Security Properties

- Security properties are reduced to looking at the impact on functional properties of looking at the system **as a whole**
- Violation of security properties affects one or more elements of the CIA triad of the system

- **Hyper-Property**: This is a set of set of traces (set of trace properties) and was introduced as important security policies cannot be expressed as properties of individual execution traces of a system
- Hyper-property amounts to looking at a system as a whole rather than individual execution traces

[N]	Set of integers 1-N i.e.[1,2,..N]
P	Set of all trace properties
W	Trace Property
w	Execution Trace



- $\phi(N) \triangleq \forall i \in [N] : \{W_i \in P_i | \forall w_i \in W_i : Power(w_i) \leq c_i\}$ where $Power(w)$ represents the power consumption of the drone and c is an arbitrary constant representing the power capacity of the drone i of the swarm
- $\phi(N) \triangleq \forall i \in [N] : \{W_i \in P_i | \forall w_{1i}, w_{2i} \in W_i : (\exists w_{3i} \in W_i : ev_{Hin}(w_{3i}) = ev_{Hin}(w_{1i}) \wedge ev_L(w_{3i}) = ev_L(w_{2i})))\}$ where ev_{Hin} represents secret events and ev_L represents public events
- $\phi(N) \triangleq \forall i \in [N] : \{W_i \in P_i | \forall w_i \in W_i : Pert(w_i, w_{orig\,i}) \leq c_i\}$ where Pert calculates the sup-norm distance between 2 traces where $w_o rig$ represents the original path trace of the drone

Tool Support for Verification of Cyber-Physical Systems

- **TLA+ Toolbox**: TLA+ is a formal specification language developed to design model, document and verify concurrent systems. The toolbox is a software tool which provides an IDE for writing and verifying TLA+ specifications
- Provides a user-friendly interface and provides support for model checking and deductive verification
- Need for support of automatic deductive verification of inductive invariants given Ordinary Differential Equations (ODEs)
- Need for support of verification of hyper-properties

Ongoing Work

- Working on contributing towards the publication of a paper on a TLA+ toolbox extension to support automatic deductive verification of inductive invariants given ODEs
- Towards publication of a paper on the support of the TLA+ toolbox for verification of hyper-properties
- Publication of a paper on a formal model of Distributed CPS for verification of security properties
- Towards the development of an automated verification technique for parameterised verification of security properties in Distributed CPS

Lancaster University

Cranfield University

UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Research Activities: RS2A

RS2-Theme A: Security in the Mission and Operational Surface

Lead: P. Angelov. Participants: A. Tsourdos, Z. Yu, Y. Li, O. Gonzalez Villarreal, A. Lopez Pellicer.

Overview

AS pose specific requirements and challenges to the detection and mitigation of cyber security risks and attacks due to their complexity and dynamic characteristics combined with the limited and unreliable network connectivity. The mission surface is the core, where the decisions and execution take place; it is dynamic and sensitive by its definition and verifiable security is of critical importance. This complicates the traditional approach that involves continual monitoring and update with patches, which links closely to the control surface below. We will develop methods and algorithms that reduce the risks and costs associated with these challenges and in turn, improve the reliability and resilience of AS.

Research activities

Autonomous Systems pose specific requirements and challenges to the detection and mitigation of cyber security risks and attacks due to their complexity and dynamic characteristics combined with the limited and unreliable network connectivity. The mission surface is the core, where the decisions and execution take place; it is dynamic and sensitive by its definition and verifiable security is of critical importance. This complicates the traditional approach that involves continual monitoring and update with patches, which links closely to the control surface below.

RS2A aims to develop methods and algorithms that reduce the risks and costs associated with these challenges and in turn, improve the reliability and resilience of AS. A thorough analysis of the existing types and classes of adversarial attacks, defence mechanisms and attack surfaces within autonomous systems has been performed. This is to identify possible scenarios and attack surfaces that may be encountered by autonomous systems in real-world scenarios such as physical adversarial patches or spoofing with malicious images that may compromise control systems.

Looking ahead...

RS2A has the following research activities planned for the next six to twelve months:

- The developed methods to detect adversarial attacks will be tested against the identified scenarios and attack surfaces to evaluate its efficiency in mitigating perturbations and new systems and scenarios will be proposed such as federated and prototype-based learning systems that are robust to adversarial attacks, and what the impact of these perturbations may have in real-world scenarios for such systems.

Research Activities: RS2A

- In order to evaluate the performance of proposed methods, collaboration with researchers developing control systems at Cranfield University will be sought to test such methods within their control systems in simulators and other environments.

Highlights & selected publications

- E. Soares, P. Angelov, and N. Suri, 'Similarity-based deep neural network to detect imperceptible adversarial attacks', IEEE Symposium Series on Computational Intelligence, IEEE SSCI, IEEE Xplore, 2022 and also presented at 2022 IEEE Symposium Series on Computational Intelligence SSCI 2022, 4th – 7th December 2022, Singapore
- Work with QinetiQ, 2ExcellGeo Ltd
- IEEE Standard on explainable AI, P2976, initiator, sponsor and WG lead initially and now a member of the WG (Prof. P. Angelov).

Research Activities: RS2B

RS2-Theme B: Securing the Control Surface

Lead: G. Inalhan. Participants: P. Angelov, A. Tsourdos, B. Yuksek.

Overview

AS relies on the ability to conduct run time adaptations of control decisions over attacks which can result from information and dynamic environment uncertainties. Specifically, in the context of learning enabled AS, it is crucial for the control system to exhibit self-aware learning in which the boundaries of “safe” state-space and the control space are tracked through their evolution. This is particularly challenging when the system is undertaking dynamic decisions within the AS mission surface.

Research activities

We have proposed an AI-based flight control system design workflow and designed an AI-based flight control system to cover the whole flight envelope of the aerial vehicle to track the given attitude commands with minimum error while providing flight safety. Nonlinear 6-degrees-of-freedom mathematical model of Boeing 737 is generated for simulation and training purposes. Classical flight control system design framework is modified to be able to integrate the learning step into the AI-based control design process.

To do so, reference models are generated based on handling quality requirements and they are utilised to calculate reference model tracking error between the actual output and reference model output. Then, a neural-network agent is trained by utilising Proximal Policy Optimization algorithm to control the attitude of the aircraft. Monte Carlo analysis is performed to evaluate the system robustness in different flight conditions. Validation of the closed-loop system dynamics is another important step for flight control systems, and it is performed by utilising frequency-domain system identification method.

In parallel with the AI-based flight control systems design study, we have been working on developing an AI-aided visual-inertial navigation system to increase the robustness of the closed-loop (i.e. control loop + feedback loop) against global positioning system (GPS) attacks such as GPS spoofing and GPS jamming which directly target the position/velocity measurements of the autonomous system. These attacks may result in catastrophic accidents in the urban area if there is no supportive system onboard.

Up to now, our achievements are; a) developing a realistic simulation environment by using Unreal Engine and Airsim in which we could get inertial, visual (RGB, depth, infrared cameras) and light detection and ranging LIDAR measurements, b) applying

Research Activities: RS2B

Research activities cont.

state-of-art visual-inertial navigation algorithms by utilising data from both simulation environment and public datasets, c) applying state-of-art segmentation algorithms in the simulation environment and public dataset

Looking ahead...

RS2B has the following research activities planned for the next six to twelve months:

- We plan to complete the robustness analysis of the AI-based flight control system against environment and aerodynamic uncertainties. A related conference paper is also planned.
- Regarding the AI-aided visual-inertial navigation system design, we plan to obtain initial results by using data from our simulation environment and public datasets. We will then perform a robustness analysis in different scenarios (i.e. weather and light conditions) for aerial and ground vehicles. Moreover, we are collaborating with RS1 to utilise Hierarchical Peer-to-Peer Federated Meta-Learning Framework for this problem to build the algorithm in a federated structure to speed up the learning process and to enable the lifelong learning.
- We also plan to continue defining the autonomous system specifications from operational safety perspective

Highlights & selected publications

- Conducting research about AI-aided intelligent mission planning and tactics development with BAE Systems.
- Collaborating with RS1 on a) developing an overview for autonomous system specifications in safety-critical applications and, b) Hierarchical Peer-to-Peer Federated Meta-Learning Framework application on AI-aided visual-inertial navigation system.

Research Activities: RS2C

RS2-Theme C: Securing the Cross-Layer Networking

Lead: W. Guo. Participants: D. Prince, P. Ciholas, Z. Wei, X. Hickman

Overview

At the Physical (PHY) level, we know digital security can be derived from both antenna beamforming (codeless defence) and deriving distributed keys from channel state information (code-based defence). The latter is particularly of interest as it can produce secure cipher keys without a common key pool or sharing keys. Yet, it must observe 3 conditions in the PHY channel, namely: (1) reciprocal to allow decentralised synchronous key generation, (2) dynamic to defence against brute force attacks, and (3) unique to avoid correlated attacks. The challenge is that the idealised conditions are often not met for ASs especially in open static spaces and airborne spaces.

Research activities

A comprehensive analysis on the threat vectors of communication surface in autonomous systems has been identified. We have shown by our works [1], [2] overleaf that the lack of channel randomness challenges the symmetric cipher key generation, but will advantage the potential attackers to crack the encrypted information. These threat vectors differ from those in cellular networks (with sufficient small-scale randomness), and thereby require novel cipher key techniques.

Looking ahead...

- A federated deep reinforcement learning based cipher key generation method is ongoing designed, which aims to extract neural network-based physical layer channel features to address the lack of randomness issue and the man-in-the-middle attacks.
- We further plan to exploit the features from autonomous control layer for legitimate cipher key generation. The control layer security (CLS) based cipher key is resistant to the attacks from physical layer, which may pave the way to secure the communication plane for adversarial autonomous scenarios.
- Work will start on developing a distributed, cross layer attack detection approach which will consider the complexities of a Networked AS swarm (spatial and mobile awareness, limited resources, etc.) which will consider optimization issues when considering mission and task planning.
- We plan to collaborate with the researchers in RS1 to see how physical layer cipher keys can cooperate with network and application layers security, e.g., the potential to provide commonality from physical layer to help build more secure differential privacy systems. In order to evaluate the performance of CLS, we plan to collaborate with researchers in RS2A-B for real UAV experiments.

Research Activities: RS2C

Highlights and selected publications.

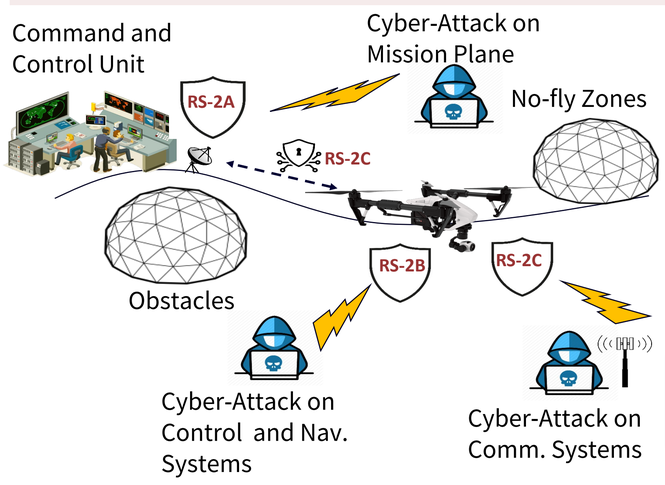
- [1] Z. Wei, W. Guo and B. Li, "A Multi-Eavesdropper Scheme Against RIS Secured LoS-Dominated Channel," IEEE Communications Letters, vol. 26, no. 6, pp. 1221-1225, 2022.
- [2] Z. Wei, L. Wang and W. Guo, "Secret Key Rate Upper-bound for Reconfigurable Intelligent Surface-combined System under Spoofing," in IEEE 96th Vehicular Technology Conference (VTC2022-Fall), 26th – 29th September 2022, London, UK.
- Federated Deep Reinforcement Learning based Physical Layer Secret Key Generation for Contaminated Channels. Z. Wei and W. Guo. Safe and Trustworthy AI Workshop, 3rd November 2022, London, UK.

RS2: Secure Operations of Trustworthy Autonomous Systems

Cranfield University, Lancaster University

Researchers: Dr. Yi Li, Dr. Zhuangkun Wei, Dr. Burak Yuksek, Dr. Oscar Villarreal, Dr. Cynthia Yu, Pierre Ciholas, Alvaro Lopez .
Investigators: Prof. Weisi Guo, Prof. Gokhan Inalhan, Prof. Plamen Angelov, Prof. Antonios Tsourdos, Prof. Dan Prince.

Secure Operation of Autonomous System



RS-2A: Exposure to cyber-physical attacks by characterizing the attack surfaces, i.e., entry points and likelihoods across mission surfaces in technology & mission-invariant manner.

RS-2B: Provide quantifiable safety and feedback to the mission surface when the limits of secure controllability are compromised

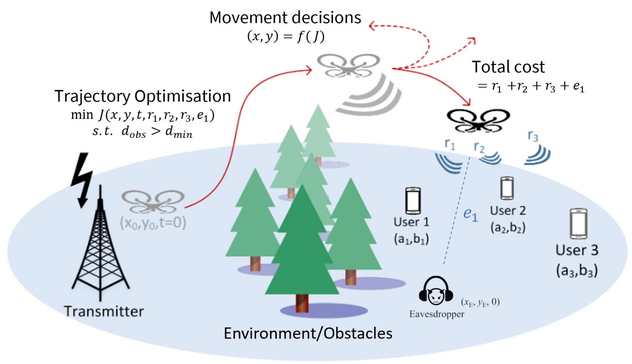
RS-2C: Provide secure communications across the different layers in the informatics plane from detection of signals to networking.

RS-2A: Securing the Mission Surface

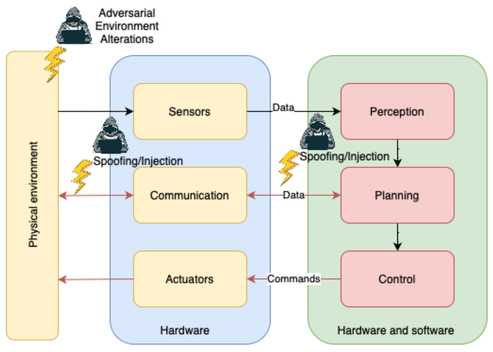
Mission Control for Secure Trustworthy Autonomous Systems requires flexible but reliable real-time optimal decision making and monitoring to handle a wide range of attacks

Methods and Focus:

- Real-Time Non-Convex Trajectory Optimization for Path Planning under constraints from control & communication
- Adaptive and Fault-Tolerant Learning-based Design for Mission Control to improve reliability of safety critical systems
- Reliable Self-Assessment under Learning-based Scenarios



Adversarial attacks



Critical Impacts

•**Perception layer**: Manipulate the sensory input of an AS, causing the system to perceive incorrect or misleading information.
•**Planning layer**: Adversarial attacks can also manipulate the AS's decision-making pro
•**Control layer**: Affect the control layer of an AS, leading to incorrect or harmful actions.

Requirements for robust to adversarial attacks systems in the context of AS:

•Able to detect attacks
•Able to react to detected attacks
•Evolve with new unknown types of attacks and situations

Physical Attacks | Digital Attacks

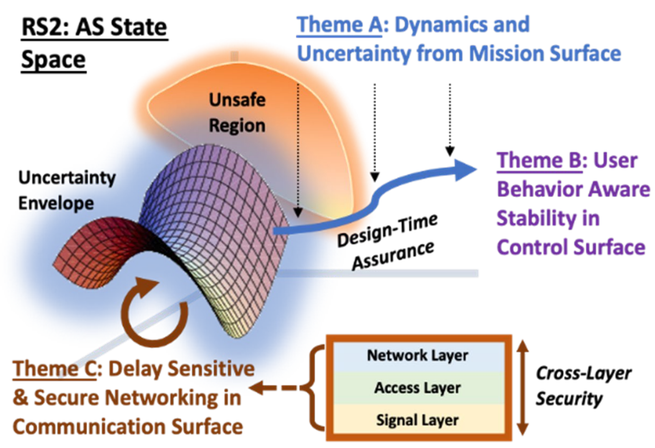
RS-2B: Securing the Control and Navigation Surface

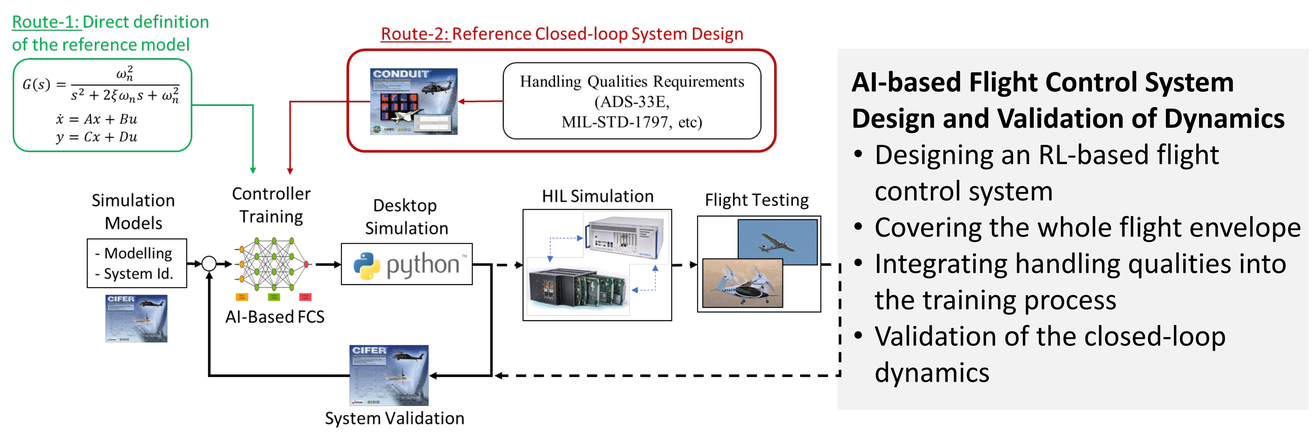
Autonomous Systems rely on the ability to conduct run time adaptations of control decisions over attacks or "perceived" attacks:

- Adversaries
- Environment uncertainties
- Degraded performance

Key Solutions for Operational Safety in Learning-Enabled Context

- AI-based Flight Control System Design and Validation of Dynamics
- AI-aided Visual Inertial Navigation for GPS-denied Environments and GPS Spoofing Detection





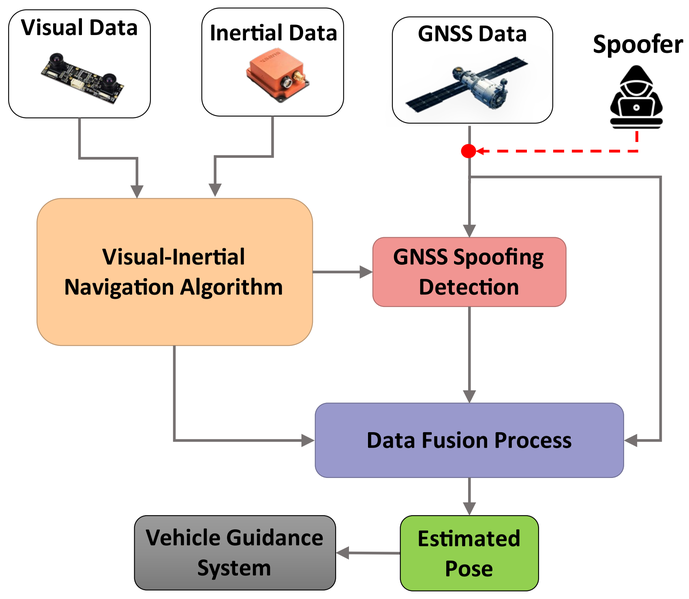
AI-based Flight Control System Design and Validation of Dynamics

- Designing an RL-based flight control system
- Covering the whole flight envelope
- Integrating handling qualities into the training process
- Validation of the closed-loop dynamics

Navigation in Extreme Adversarial Environment

AI-aided Visual Inertial Navigation for GPS-denied Environments and GPS Spoofing Detection

- Designing AI-aided Visual-Inertial navigation system to support the GNSS in the presence of spoofing attacks.
- Combining the AI-based solutions with classical filter-based approach
- Improving the pose estimation performance in austere environment
- **Case studies;**
  - Civil: Urban air mobility
  - Military: Perceptional intelligence in austere environments



RS-2C: Securing the Communication Surface

Physical & Control Layer Security

To secure the communication surfaces of AS, current cryptography and physical layer security (PLS) both have some severe security threats, which motivates the design of control layer security (CLS) that is specific for AS.

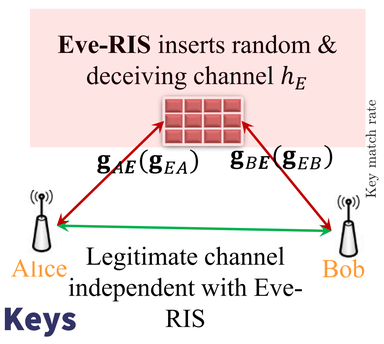
Cryptography

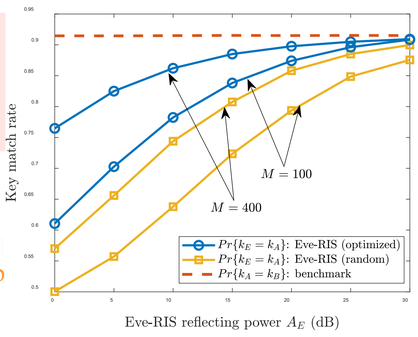
uses common key pool for cipher key generation, but has following issues:

- Complex key generation & management & distribution
- No secrecy guaranteed against post-quantum computing
- High computational complexity & latency

PLS

generates cipher keys via the reciprocal channel information, but has man-in-the-middle attack threats:

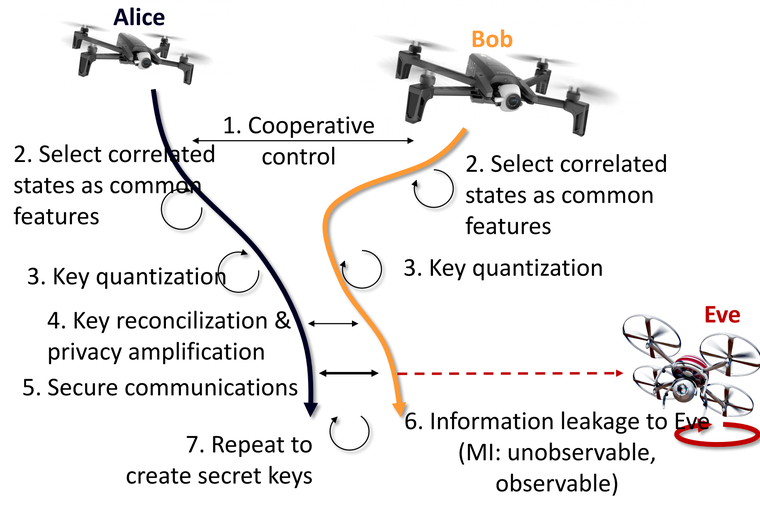


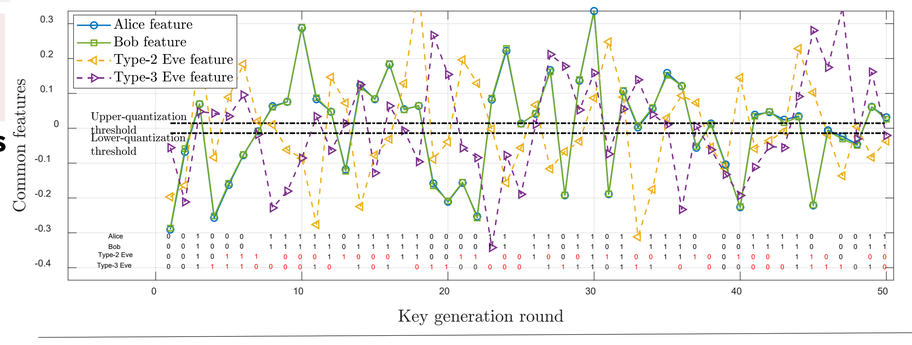


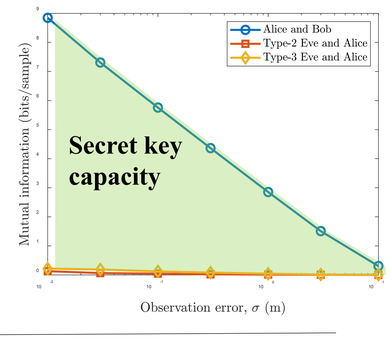
Designed Control Layer Cipher Keys

Legitimate Alice and Bob (two UAVs) create correlated but unobservable states (e.g., yaw angles), via cooperative control, and use these correlated states for cipher key generation.

Results show that by properly designing the cooperative control algorithm, UAV Alice and UAV Bob can have random but highly correlated states for cipher key generation, which prevent attackers from eavesdropping.





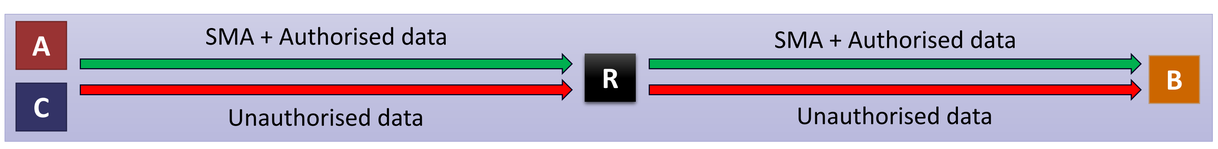


Data Link, Network, and Transport Layer Security

To secure communication on higher layers (Data Link, Transport, and Session), and potentially create a secure and trusted network within an untrusted network, we use cryptography assuming a pre-shared secret and a Single Message Authentication protocol of our creation.

Timestamp	User ID	Entropy	Access	MAC
Clear		Encrypted (KD/OTP)		Clear

The packet of the AS is intercepted before being sent, and we perform the authentication first before forwarding the AS packets. If the authentication succeeds, a single communication is allowed to get through (for stateful protocols such as TCP), whereas for stateless protocols (e.g. UDP) other solutions are available (e.g. merge authentication and data, of provide a hash of the expected payload). This solved the infamous "NAT problem" as it is known in the literature.



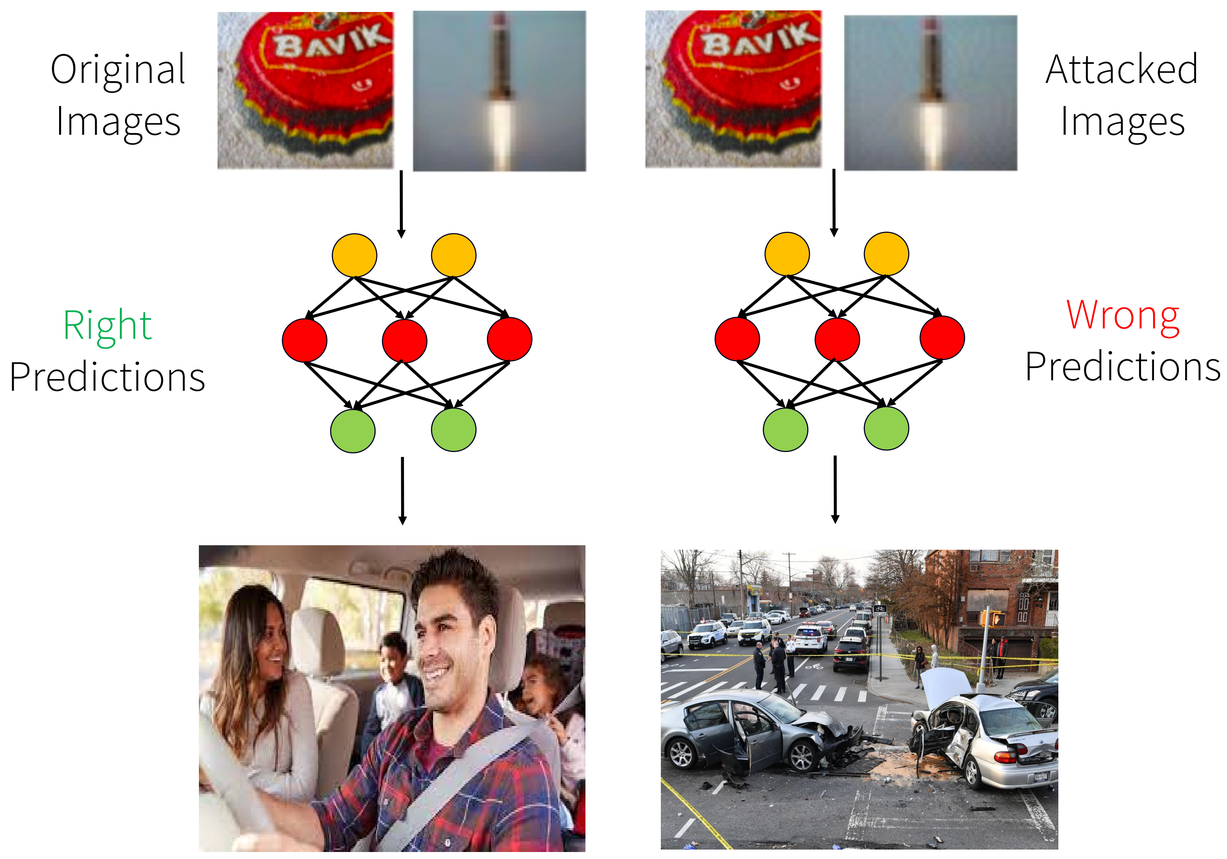
This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Domain Generalization and Feature Fusion for Cross-domain Imperceptible Adversarial Attack Detection

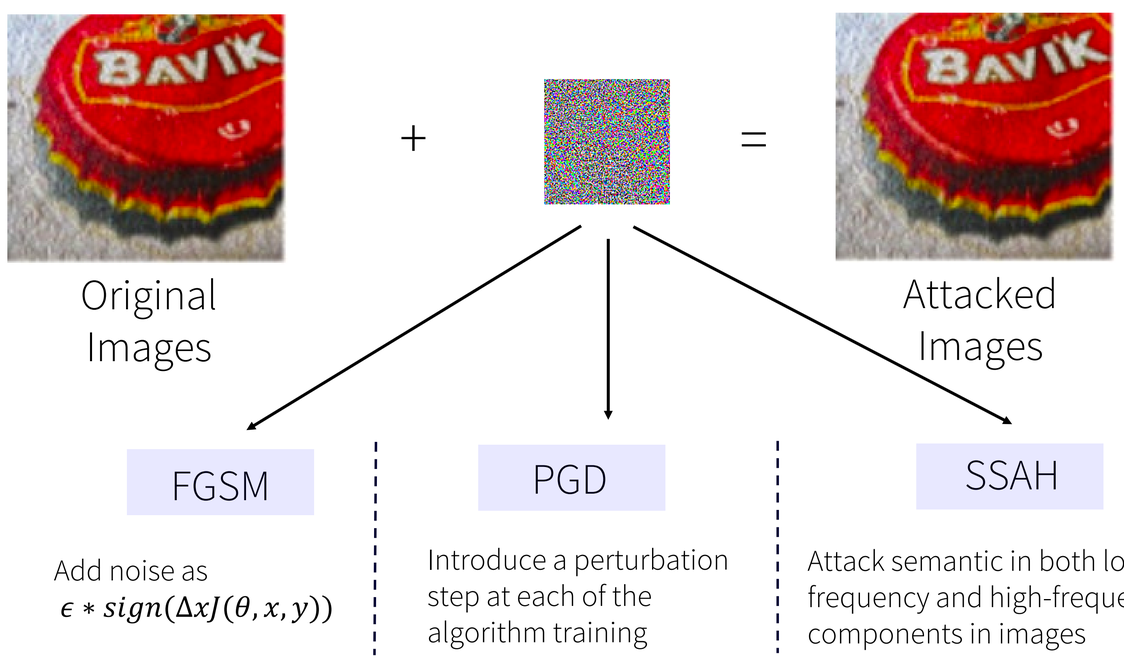
Lancaster University

Researcher: Dr. Yi Li
Investigators: Prof. Plamen Angelov, Prof. Neeraj Suri

Background: Imperceptible Adversarial Attack Detection

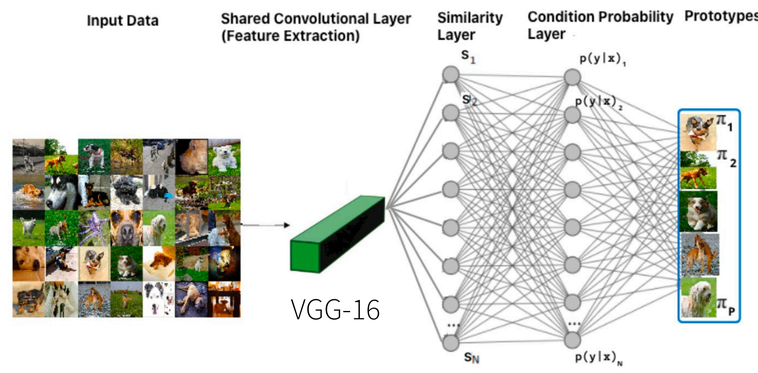


Attacks



Learning-Based Detection Methods: State-of-the-art

Sim-DNN



Pros:

- These methods provide excellent results for various attacks.
- These methods require few manual-engineering

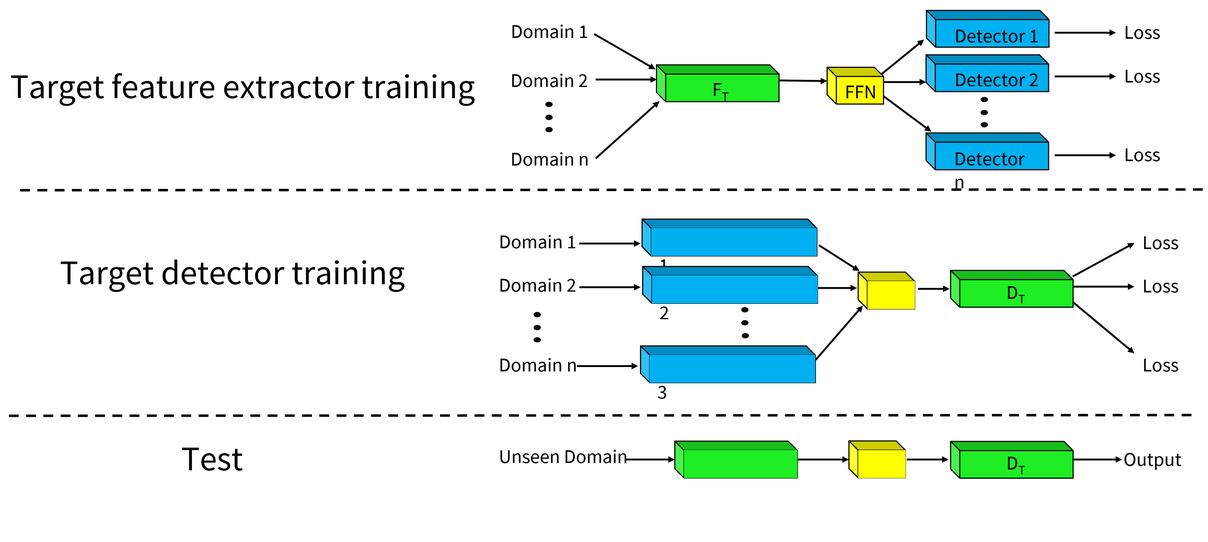
Cons:

- Weak adaptability and transferability to new domains, e.g., attacks or datasets.
- Slow training due to large model scales, particularly for the feature extractor (VGG-16).

Engineering and Physical Sciences Research Council

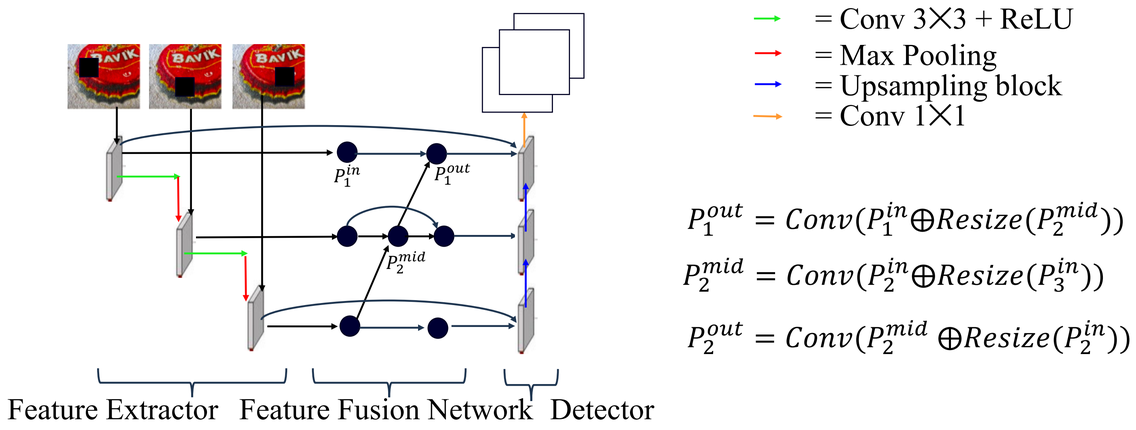
TAS-S

Domain Generalization Framework



- The feature extractor or detector is trained with a partner who is well tuned for different domains.
- In the test stage, the trained target feature extractor and detector are combined with the FFN to detect attacks in unseen domains.

Feature Fusion Network (FFN)



$$P_1^{out} = Conv(P_1^{in} \oplus Resize(P_2^{mid}))$$
$$P_2^{mid} = Conv(P_2^{in} \oplus Resize(P_3^{in}))$$
$$P_2^{out} = Conv(P_2^{mid} \oplus Resize(P_2^{in}))$$

Experimental Results

- Same attack in training and test
- Different datasets in training and test
- 10k images in ImageNet-R as the test dataset
- Training datasets: CIFAR-10, CIFAR-100, ILSVRC
- 50k images from each dataset for training

Attack Detection Performance Comparisons

	Computational Complexity		Detection Ratio (%)		
Method	Para. (M)	Time (s)	FGSM	PGD	SSAH
	37.9		55.1 ± 1.8	59.2 ± 2.3	48.8 ± 2.5
Epi-FCR	62.7	728.4	57.3 ± 1.4	60.0 ± 1.6	49.4 ± 0.7
Adversarial	8.2	102.1	57.8 ± 1.3	61.1 ± 1.1	49.5 ± 1.7
L-RED					53.1 ± 1.5
Sim-DNN	134.9	1291.6	64.5 ± 1.6	66.9 ± 1.9	
DGAD (3)	**2.1**	**99.6**	67.3 ± 0.9	69.4 ± 1.1	65.6 ± 0.8
DGAD (4)	4.8	141.7	69.5 ± 0.7	72.2 ± 1.0	69.9 ± 0.5
DGAD (5)	6.9	168.0	**75.0 ± 0.4**	**76.3 ± 0.5**	**72.5 ± 0.5**

- Different attack in training and test
- Different datasets in training and test
- 10k images in ImageNet-R as the test dataset

Attack Detection Performance Comparisons

	Detection Ratio (%)		
Method	FGSM	PGD	SSAH
MetaQDA	50.4 ± 2.0	55.5 ± 2.1	43.7 ± 2.9
Epi-FCR	56.9 ± 1.6	59.4 ± 1.6	49.1 ± 0.8
Adversarial	53.2 ± 1.9	57.6 ± 1.4	45.5 ± 1.8
L-RED	56.1 ± 1.4	58.8 ± 1.5	48.2 ± 2.1
Sim-DNN	60.8 ± 1.7	63.3 ± 2.4	55.2 ± 1.5
DGAD (3)	65.8 ± 0.8	68.1 ± 1.3	64.0 ± 1.0
DGAD (4)	68.9 ± 0.7	71.0 ± 1.3	69.1 ± 0.7
DGAD (5)	**73.8 ± 0.6**	**73.2 ± 0.9**	**69.5 ± 0.7**

Ongoing and Future Works

- Visualization results of the proposed algorithm will be completed.
- Adaptability and transferability will be evaluated in real-world pictures, e.g., infrastructure.
- Ablation study of the proposed algorithm will be provided.

Lancaster University

Cranfield University

UKRI Trustworthy Autonomous Systems Hub

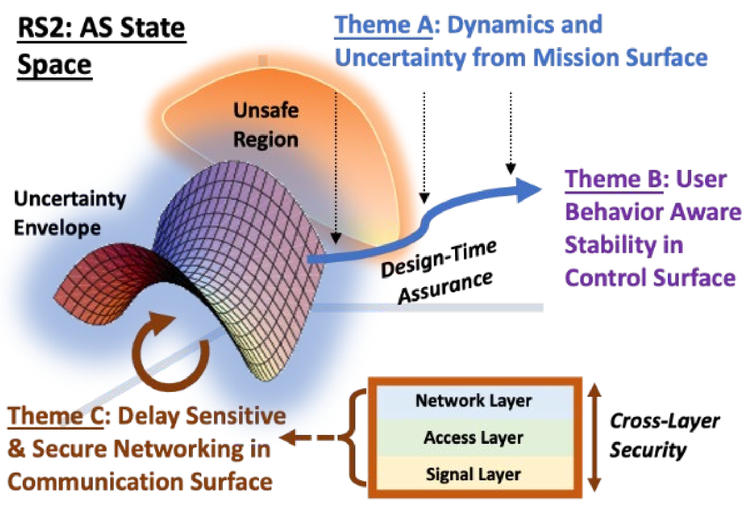
This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Securing the Control and Navigation Surfaces

Cranfield University

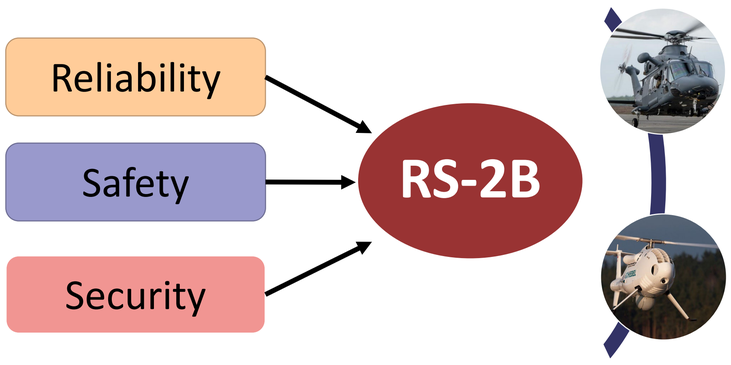
Researcher: Dr. Burak Yuksek
Investigator: Prof. Gokhan Inalhan

1. Role of the RS-2B



Ability of runtime adaptations of control and navigation systems over attacks or "perceived" attacks:

- Adversaries
- Environment uncertainties
- Degraded performance



Design and Dynamical V&V of AI-based Flight Control Systems

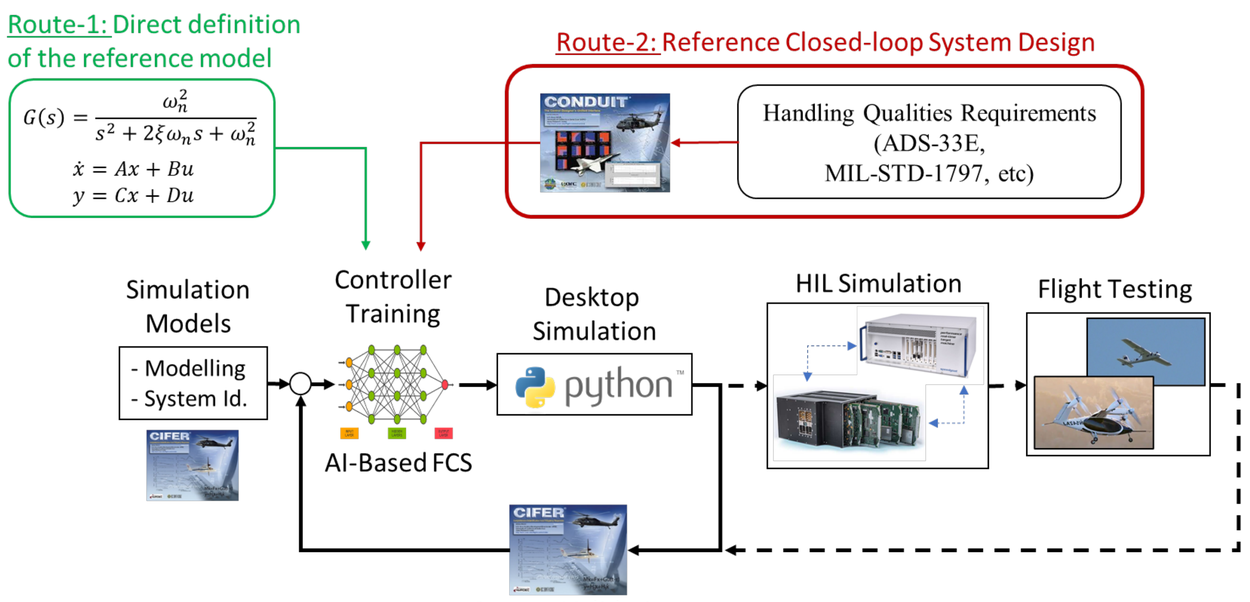
AI-aided Visual Inertial Navigation Systems

1. AI-Based Flight Control System Design

Aim of this Study:

- Designing an RL-based flight control system
  - Covering the whole flight envelope
  - Integrating handling qualities into the training process
  - Validation of the closed-loop dynamics

Design Methodology Overview



Validation of the Closed-loop System Dynamics in Simulation Environment

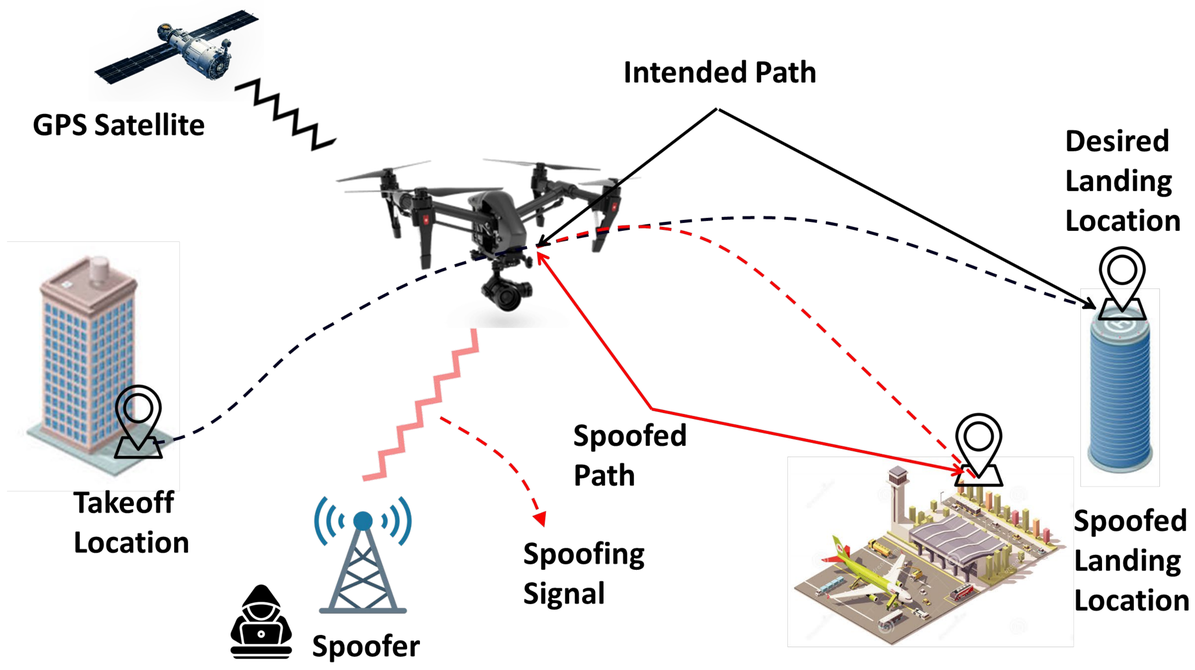
Summary of Dynamical Validation Tests in Simulation Environment

		Roll Axis			Pitch Axis		
		AI FCS	Ref Model	Req.	AI FCS	Ref Model	Req.
Broken-loop Analysis	0dB Crossover Freq (rad/s)	4.556	2.165	> 2 rad/s	2.9176	3.0598	> 2rad/s
	PM (deg)	40.634	46.866	> 45 deg	44.1568	45.636	> 45 deg
	GM (dB)	19.675	13.880	≥ 6 dB	23.2805	10.828	≥ 6dB
Disturbance Rejection	DRP (dB)	3.939	4.435	< 5 dB	3.8222	4.631	< 5 dB
	DRB (rad/s)	1.906	0.820	> 1 rad/s	1.4876	0.854	> 1 rad/s

Handling Quality Levels Level 1 Level 2 Level 3

PM: Phase Margin, GM: Gain Margin, BW: Bandwidth, DRP: Disturbance rejection peak, DRB: Disturbance rejection bandwith, Req.: Requirement

2. AI-aided Visual Inertial Navigation (VIN) for GPS-denied Environments and GPS Spoofing Detection

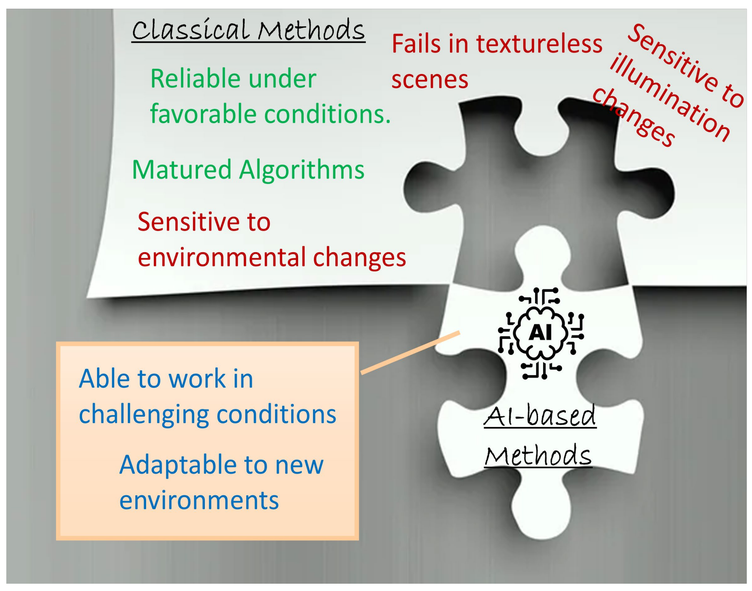


Operations in Urban Airspace

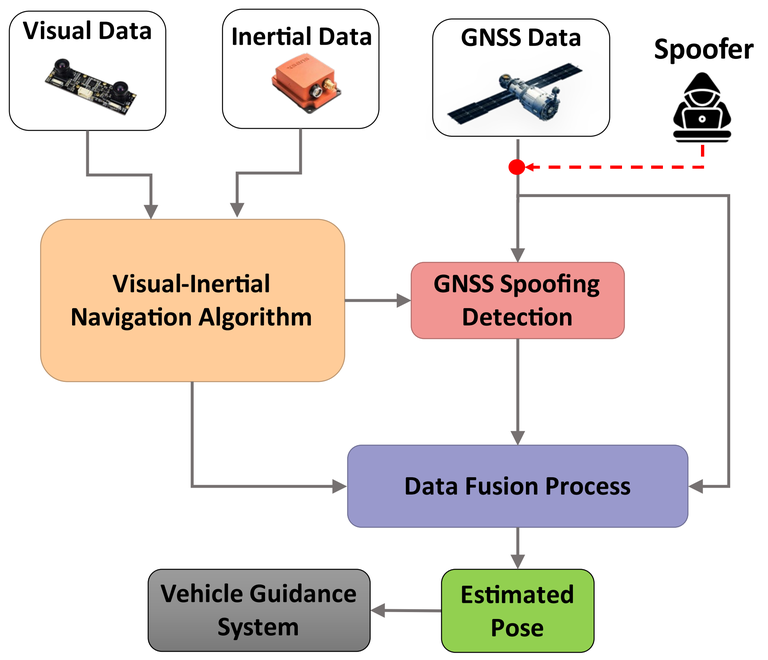
- Require high level of safety
- GNSS is one of the most vulnerable system against cyber-attacks such as jamming and spoofing
- Spoofing attacks are more harmful and difficult to detect
- GNSS system should be supported by utilising multi-sensor pose estimation algorithms not only to detect the attacks but also to provide safety for the vehicle.

Research Proposal

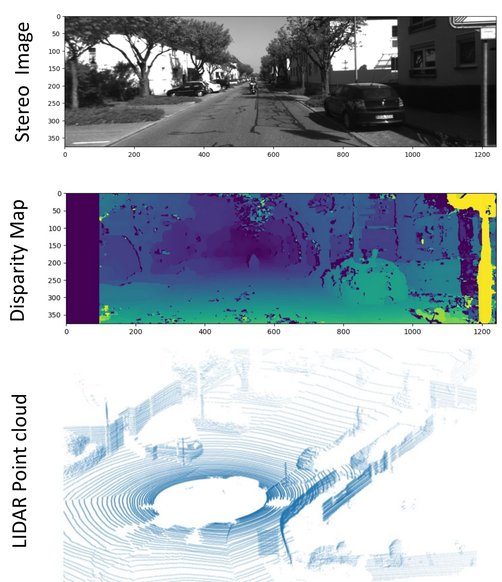
- Designing AI-aided Visual-Inertial navigation system to support the GNSS in the presence of spoofing attacks.
- Combining the AI-based solutions with classical filter-based approach
- Improving pose estimation performance in austere environments



AI-aided VIN System and GPS-Spoofing Detection Overview



Initial Studies by using KITTI Dataset (for ground vehicles)



3. Conclusions

AI-based FCS Design

It is shown that it is possible to integrate handling quality requirements into training process of the AI-based flight control system and validate it by utilizing frequency-domain system identification method.

AI-aided Visual-Inertial Navigation System Design

One of the most dangerous cyber-attacks on autonomous systems in urban environment is GNSS spoofing attack. It is required to support it by utilizing visual-inertial navigation solutions. AI has a significant role to improve the navigation solution accuracy in austere environments and to make the GNSS spoofing detection system more reliable.

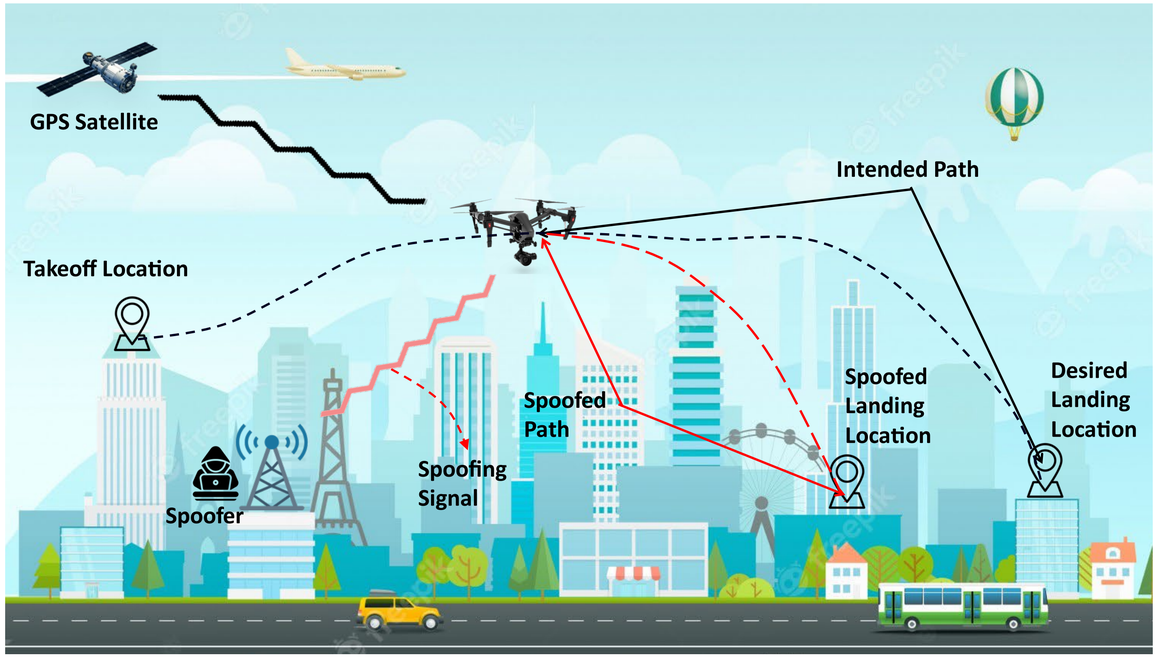
This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Federated Meta Learning for UAV Visual Navigation in Urban Airspace in the Presence of GPS-Spoofing Attacks

Cranfield University & Lancaster University

Researchers: Dr. Burak Yuksek, Dr. Zhengxin Yu
Investigators: Prof. Gokhan Inalhan, Prof. Neeraj Suri

Visual Navigation for Autonomous Vehicles



Operations in Urban Airspace

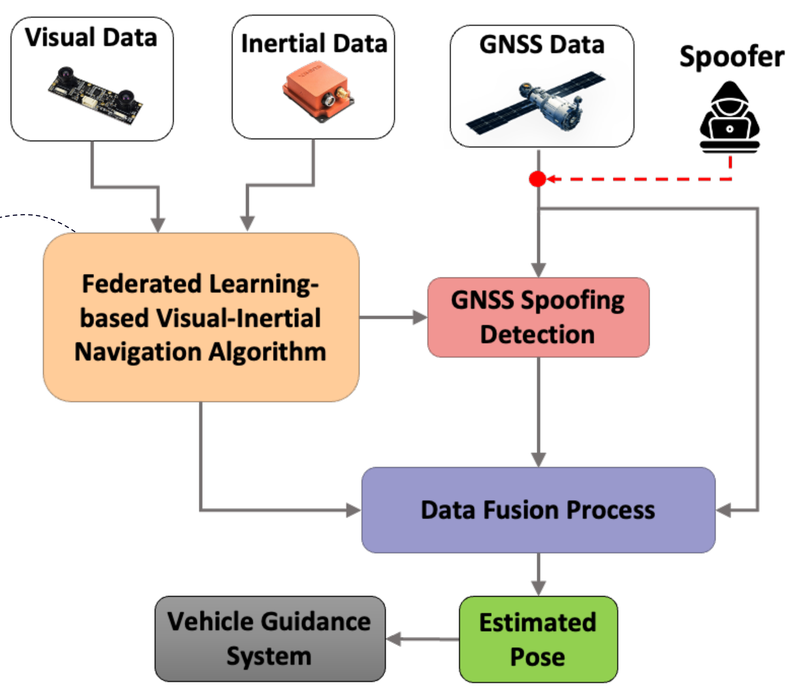
- Require high level of safety
- GNSS is one of the most vulnerable system against cyber-attacks such as jamming and spoofing
- Spoofing attacks are more harmful and difficult to detect
- Measurement errors such as multi-path error should be compensated for high positioning accuracy
- GNSS system should be supported by utilising multi-sensor pose estimation algorithms not only to detect the attacks but also to provide safety for the vehicle.

Design Goals

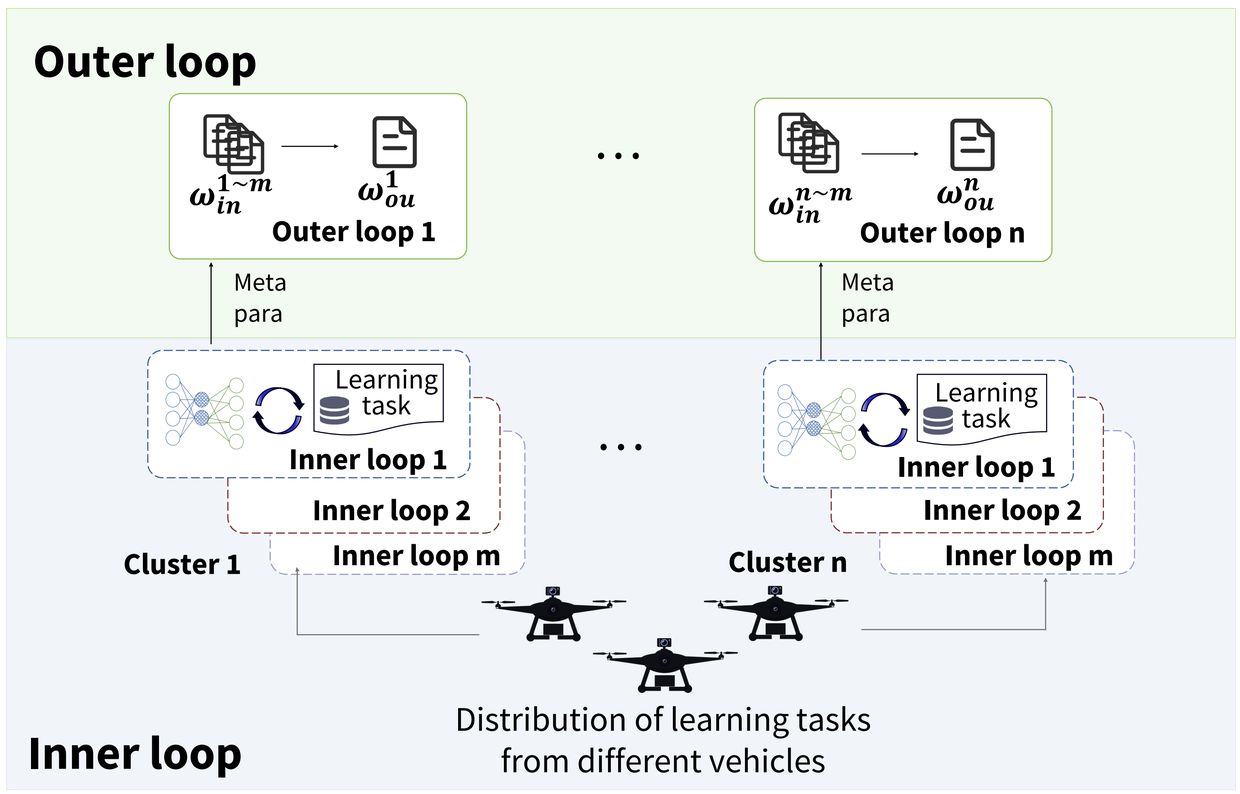
- GNSS-spoofing attack detection
- Increasing pose estimation accuracy
- Generalizability and adaptation for new environments
- Lifelong learning ability

Federated Learning-based Visual Odometry Framework

- Combining the AI-based solutions with classical filter-based approach
- Utilising federated learning framework to improve pose estimation accuracy.
- Aggregating models trained in different environments and conditions .

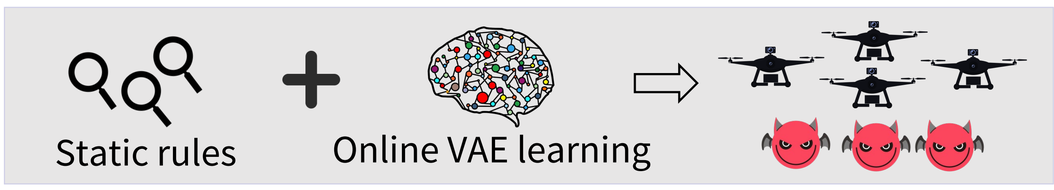


Adaptive and Robust Federated Meta Learning



A P2P federated learning + meta-learning for navigation

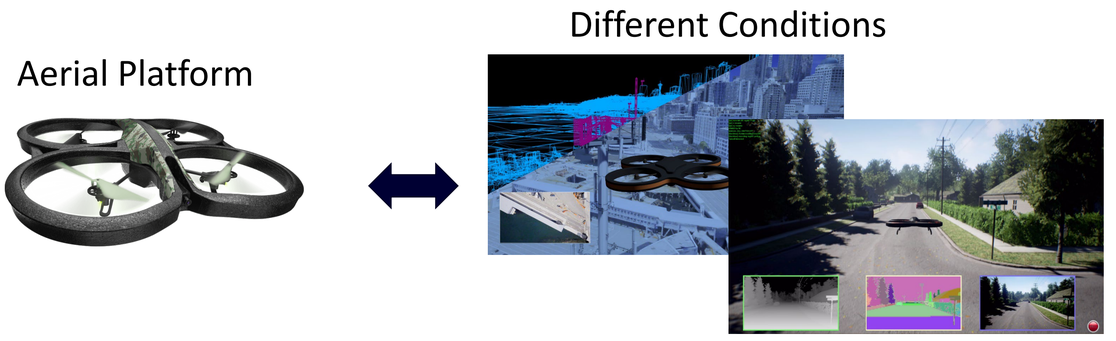
- **An adaptive meta-learning architecture** is proposed to adapt to new environments and enable vehicles to have the lifelong learning capability.
  - Inner loop:
    - Train a task-specific model based on local data
  - Outer loop:
    - Extract common features from similar tasks
    - Optimize meta-model adaptability of similar tasks
- **A robust-by-design federated meta-learning architecture** is developed to adaptively defend against a range of adversarial attacks.
  - A composite rule-based and learning-based detection method to effectively identify adversarial vehicles via ranking domain and low-dimensional embeddings.
  - An adaptive model aggregation method aggregate the global model by considering the degree of similarity between the meta-model and calculated mean model to resilience attacks.



Detection Models — Outer loops

Simulation Framework

- Unreal Engine and AirSim
- Nonlinear dynamical model for aerial vehicles
- Realistic sensor models (IMU, GNSS, LIDAR)
- Photorealistic Camera Data Monocular and Stereo
  - Gray scale and RGB



Ongoing and Future Works

- Implementation of the proposed algorithm will be completed.
- Adaptability and transferability will be evaluated in outdoor environments for different weather and light conditions.

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Threat Analysis of current Physical Layer Security on Communication Surfaces of Autonomous Systems

Cranfield University

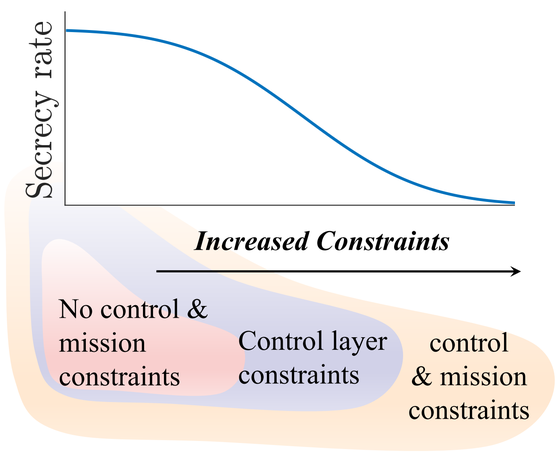
Researcher: Dr. Zhuangkun Wei

Investigator: Prof. Weisi Guo

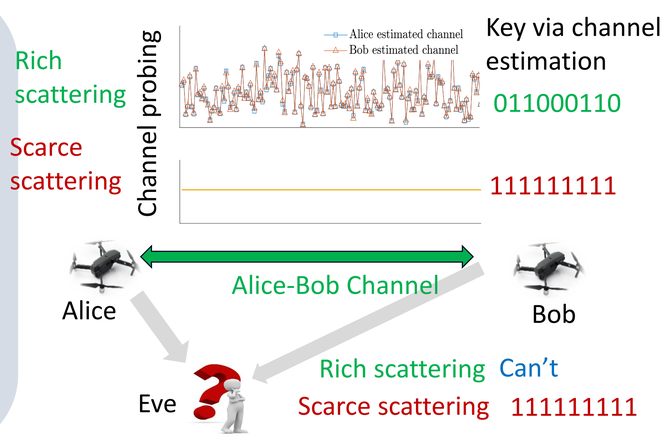
Introduction

Communications of autonomous systems are vulnerable to attacks and eavesdropping, due to broadcasting communication nature and the lack of randomness of communication channels

Key-Less Physical Layer Security (key-less PLS):
maximize secrecy rate or signal-to-interference-noise-ratio (SINR), by optimizing trajectory, beamforming, IRS phase.
Advantage: key-less, easy deployment
Disadvantage: no solution guarantee when combined with mission & control layers objectives & constraints



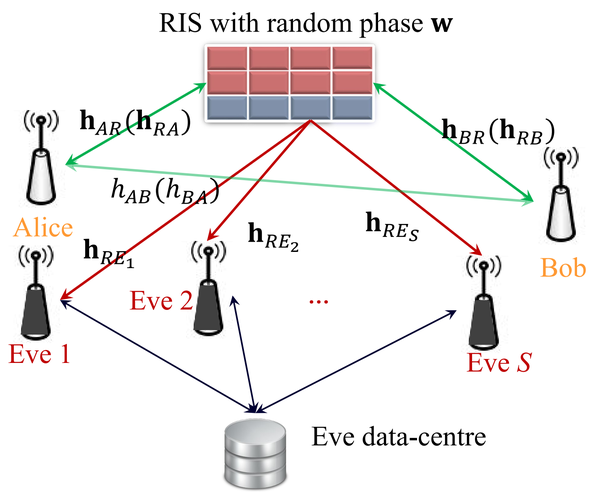
Physical Layer Secret Key Generation (PL-SKG):
Generate shared secret sky via the reciprocal small-scale channel randomness.
Advantages: detached from mission & control layer optimization
Disadvantages: requires sufficient small-scale scattering & randomness



1. Cooperative Passive Eavesdropping Threat

Reconfigurable intelligent surface (RIS) is a promising technology to secure the LoS dominated low-entropy channels, by inducing randomness via IRS phases

However, the RIS-induced randomness is also contained in the Eves' received signals, which enables the estimation of the legitimate channel by multiple & cooperative Eves.



Theory of Multi-Eve Design

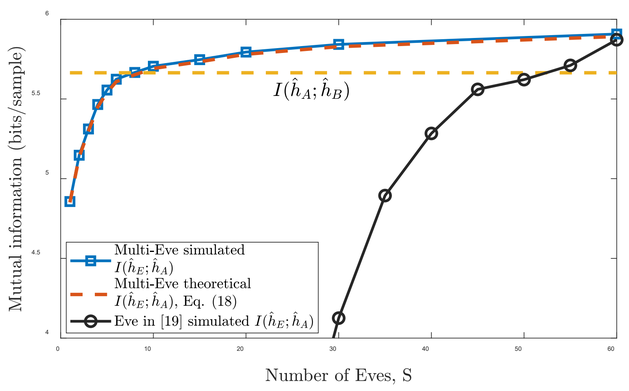
Consider *S* Eves, each Eve's received signals are:

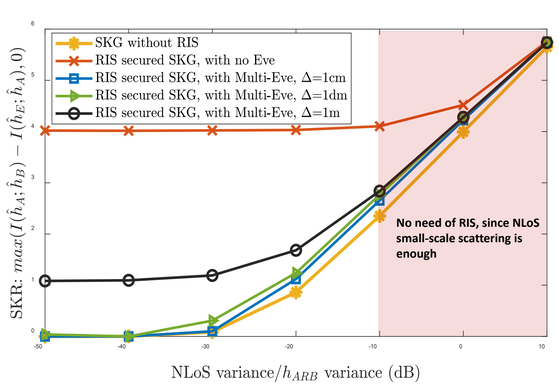
$$\mathbf{z}_s^{(odd)} = (h_{AE_s} + \mathbf{h}_{RE_s} \cdot diag(\mathbf{w}) \cdot \mathbf{h}_{AR}) \cdot \mathbf{u}_A + \boldsymbol{\varepsilon}_s^{(odd)}$$
$$\mathbf{z}_s^{(even)} = (h_{BE_s} + \mathbf{h}_{RE_s} \cdot diag(\mathbf{w}) \cdot \mathbf{h}_{BR}) \cdot \mathbf{u}_B + \boldsymbol{\varepsilon}_s^{(even)}$$

The deployment of *S* Eves is to ensure the conditional entropy of legitimate channel on *S* Eves' received equals 0, which suggests a successful estimation of the legitimate channel from Eves.

$$H\left(h_{BA} + h_{ARB} | \mathbf{z}_1^{(odd)}, \mathbf{z}_1^{(even)}, \cdots, \mathbf{z}_S^{(odd)}, \mathbf{z}_S^{(even)}\right) \overset{(a)}{\approx} H\left(\mathbf{h}_{RA} diag(\mathbf{h}_{BR}) \cdot \mathbf{w}; \begin{bmatrix} \mathbf{H}_{RE} \cdot diag(\mathbf{h}_{AR}) \\ \mathbf{H}_{RE} \cdot diag(\mathbf{h}_{BR}) \end{bmatrix} \cdot \mathbf{w}\right) \overset{(b)}{=} 0$$

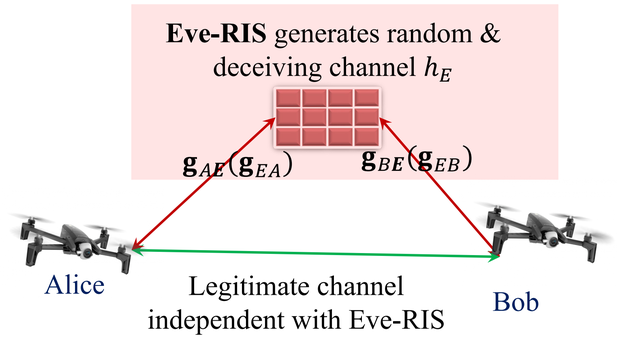
Results of Cooperative Eve Design



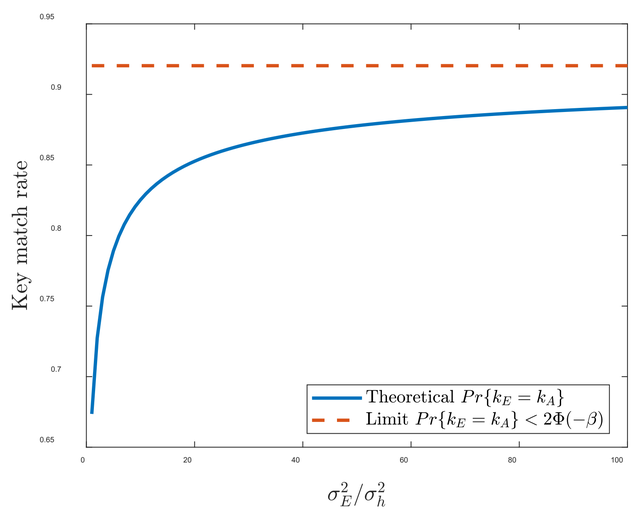


2. Eve-RIS: Concealed Man-in-the-middle Attack

With the advancement of RIS, an adversarial RIS can be used to generate and insert a deceiving channel to the legitimate channel, and then derive the legitimate secret keys. This is a more concealed way of man-in-the-middle attack, since RIS is naturally resistant to countermeasures for untrusted relay.



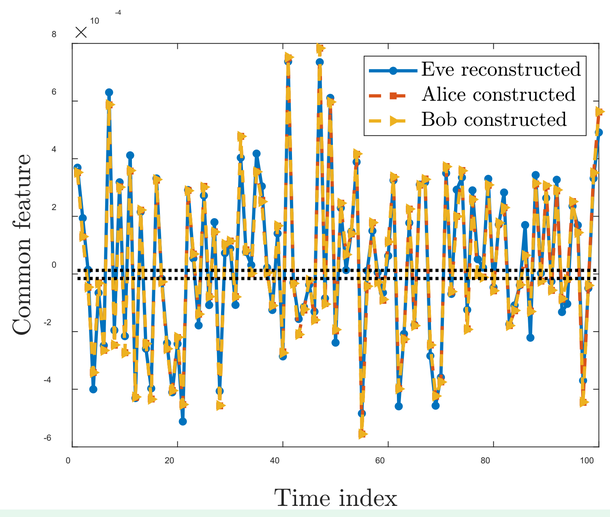
Theory of Eve created Channel Randomness

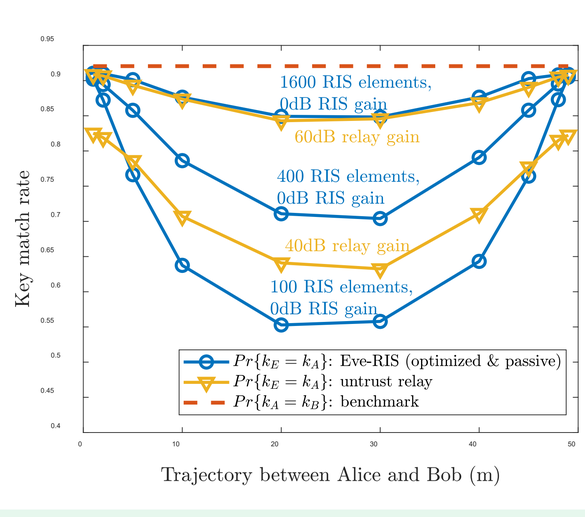


$$Pr\{k_A = k_E\} = \sqrt{\frac{2}{\pi}} \int_{\beta}^{+\infty} \Phi\left(-\beta \sqrt{\frac{\sigma_E^2}{\sigma_h^2} + 1} + \frac{\sigma_E}{\sigma_h}\zeta\right) \exp\left(-\frac{\zeta^2}{2}\right) d\zeta$$

With the increase of the variance of Eve-RIS's inserted channel, σ_E^2, the theoretical key match rate between Eve-RIS and legitimate user increases drastically, indicating its potential of stealing the cipher keys

Results of Eve-RIS



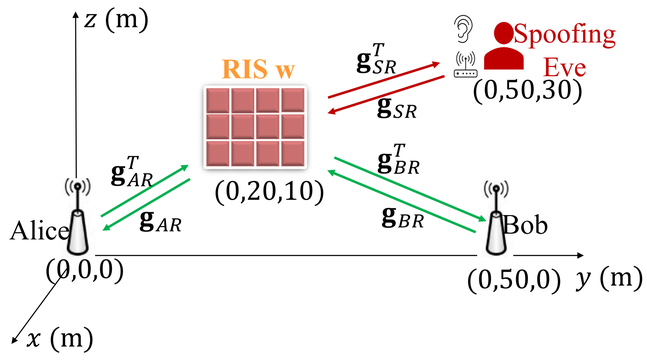


3. Spoofing: with friendly or adversarial RIS

Sketch of Pilot Spoofing

A spoofing Eve aims to pretend as Alice, by sending an amplified Alice's pilot sequence by ρ, simultaneously in the Alice's sending time-slot

$$SKR_L \triangleq \max\left\{I\left(\hat{h}_A; \hat{h}_B\right) - I\left(\hat{h}_S; \hat{h}_B\right), 0\right\}$$
$$SKR_S \triangleq I\left(\hat{h}_S; \hat{h}_B\right)$$



Upper-bound of Legitimate SKR

Theorem 2: When $\sigma_\epsilon^2 \to 0$ (i.e., with high receiving signal-to-noise ratio, SNR), the legitimate SKR has an upper-bound as:

$$SKR_L < \max\left\{0.5 \log_2 \frac{1}{\rho^2} \lambda_{max}\left((\mathbf{U}_{SB}^{-1})^H \mathbf{R}_{AB} \mathbf{U}_{SB}^{-1}\right), 0\right\},$$

where $\lambda_{max}(\cdot)$ represents the maximal eigenvalue of a matrix. $\mathbf{U}_{SB} \triangleq \mathbf{\Lambda}_{SB}^{0.5} \mathbf{\Gamma}_{SB}^H$, with the eigen-decomposition of $\mathbf{R}_{SB}$, i.e., $\mathbf{R}_{SB} = \mathbf{\Gamma}_{SB} \mathbf{\Lambda}_{SB} \mathbf{\Gamma}_{SB}^H$.

Upper-bound of Spoofing SKR

Theorem 3: The spoofing SKR is bounded by:

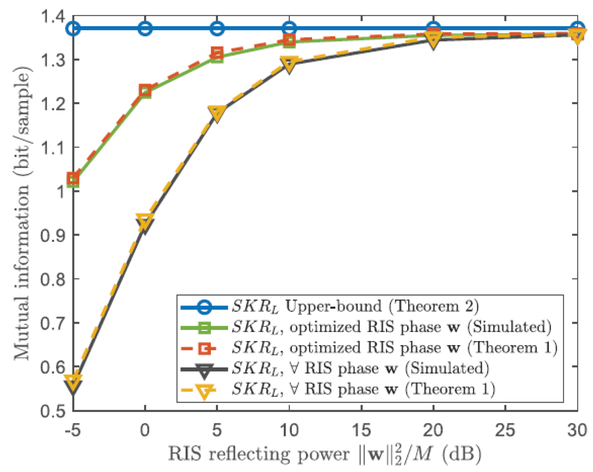
$$SKR_S < 0.5 \log_2\left(1 + \rho^2 \lambda_{max}\left((\mathbf{U}_{AB}^{-1})^H \mathbf{R}_{SB} \mathbf{U}_{AB}^{-1}\right)\right), \quad (12)$$

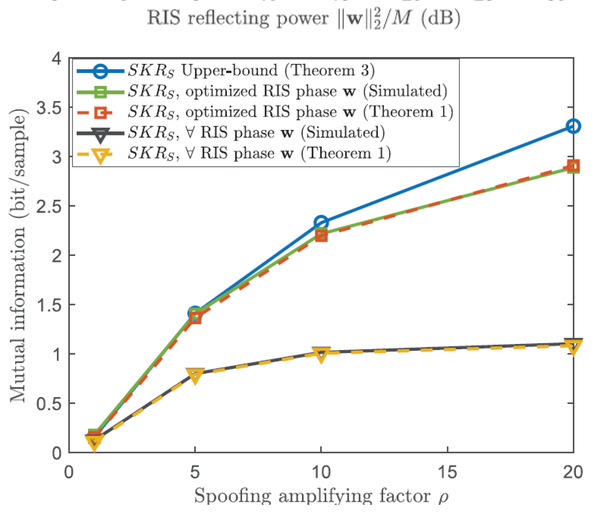
where $\mathbf{U}_{AB} \triangleq \mathbf{\Lambda}_{AB}^{0.5} \mathbf{\Gamma}_{AB}^H$, with the eigen-decomposition of $\mathbf{R}_{AB}$, i.e., $\mathbf{R}_{AB} = \mathbf{\Gamma}_{AB} \mathbf{\Lambda}_{AB} \mathbf{\Gamma}_{AB}^H$.

One sub-optimal solution

$$(\mathbf{R}_{SB} - \lambda_{max}(\mathbf{R}_{SB}) \cdot \mathbf{I}_M) \cdot \mathbf{w}_{\text{s-opt}} = \mathbf{0}$$

Results show that RIS can help little against pilot spoofing in autonomous systems, but can be used to improve the spoofing if used by adversarial users





This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Control Layer Secret Key Generations for Autonomous Systems

Cranfield University

Researchers: Dr. Zhuangkun Wei, Dr. Oscar J. Gonzalez V.

Investigators: Prof. Weisi Guo, Prof. Antonio Tsourdos

Introduction

Current strategies to secure the communication surfaces of autonomous systems include cryptography and physical layer security (PLS). However, both have some severe security issues (shown in the following), which motivates the design of control layer security (CLS) that is specific for autonomous systems.

Cryptography

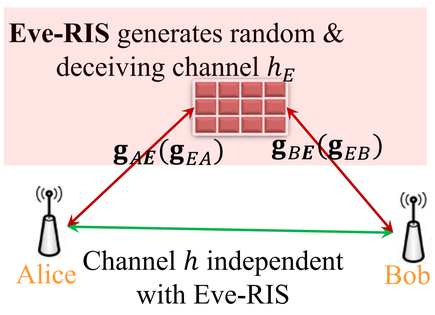
uses common key pool for cipher key generation, but has following issues:

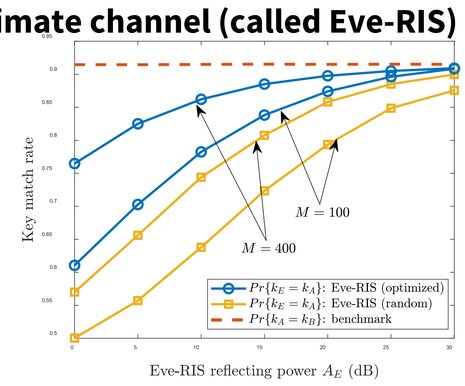
- Complex key generation & management & distribution
- No secrecy guaranteed against post-quantum computing
- High computational complexity & latency

Physical Layer Security

generates shared secret keys via the reciprocal small-scale channel randomness of Alice and Bob, however, has following attack threats:

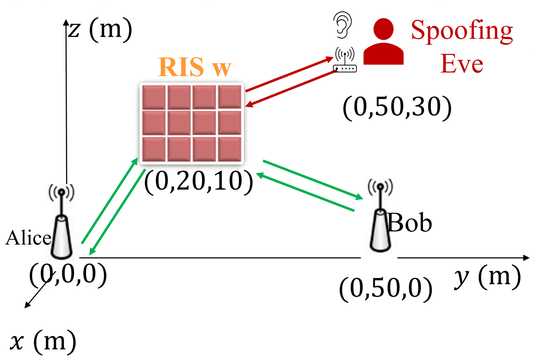
(1) When an adversarial reconfigurable intelligent surfaces (RIS) inserts a deceiving channel into the legitimate channel (called Eve-RIS)

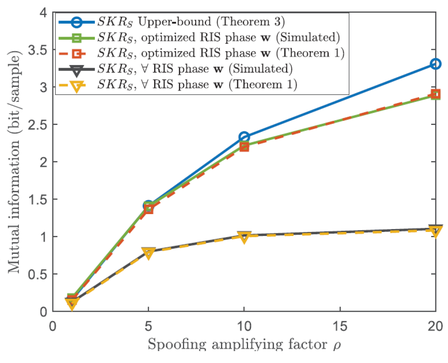




Results show that Eve-RIS can have high key match rate with legitimate users, therefore able to derive the cipher keys

(2) A spoofing Eve assisted by an adversarial RIS



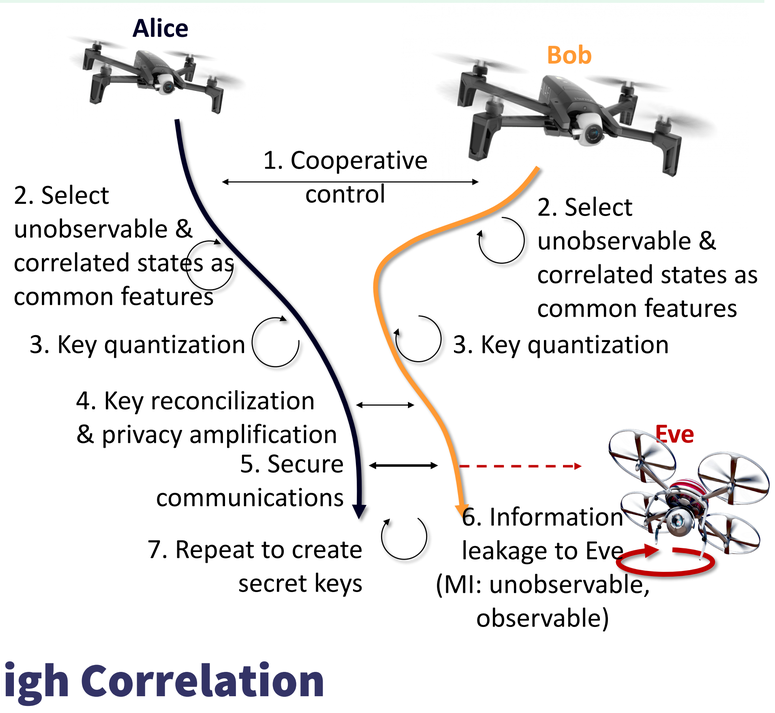


Results show that adversarial RIS can be used to improve the spoofing if used by adversarial users

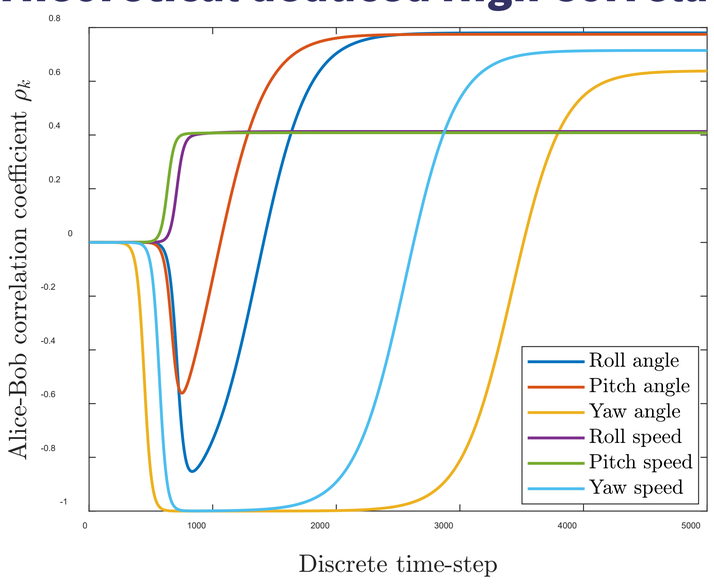
1. Concept & Theory of Control Layer Security

Legitimate Alice and Bob (two UAVs) create correlated but unobservable states (e.g., yaw angles), via cooperative control, and use these correlated states for cipher key generation.

$$\mathbf{x}_k^{(i)} = \mathbf{A} \cdot \mathbf{x}_{k-1}^{(i)} + \mathbf{B} \cdot \mathbf{u}_{k-1}^{(i)}$$
$$\mathbf{y}_k^{(i)} = \mathbf{C} \cdot \mathbf{x}_k^{(i)} + \boldsymbol{\varepsilon}_k^{(i)}$$
$$i \in \{Alice, Bob\}$$



Theoretical deduced High Correlation



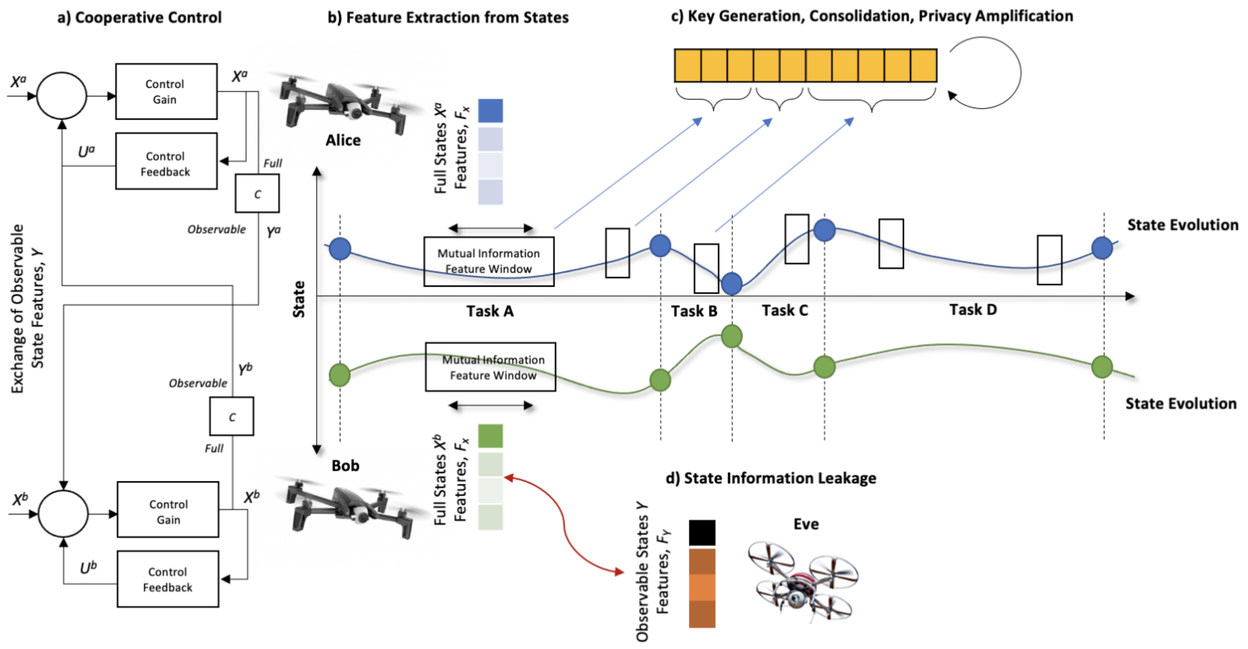
An appropriate cooperative control design can make the correlation between the states of two UAVs approach to ± 1, rendering the potential to use these highly correlated states for cipher key generation, which avoids suffering from the aforementioned threats of cryptography and PLS

2. Difference from Physical Layer Security

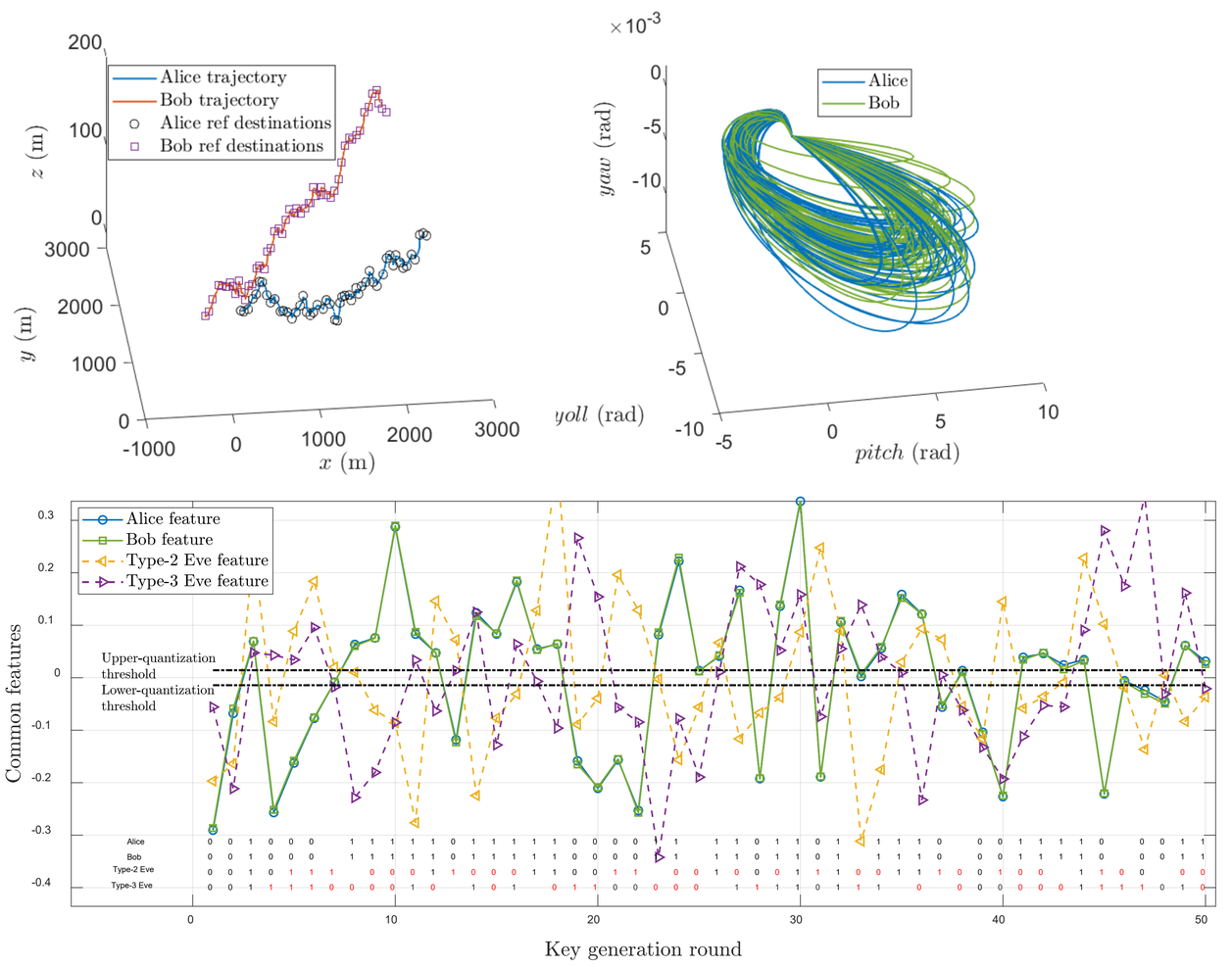
	Prerequisites	Available channel noise by jamming, pilot spoofing	Available positioning error
CLS (proposed)	Cooperative control, multiple to one map from unobservable to observable states	Not affected by channel attacks	cm-m level, to ensure selected states with correlation >0.8
PLS	Channel reciprocity, randomness	<-10dB s.t. correlation coefficient >0.8	Not affected by position observation error

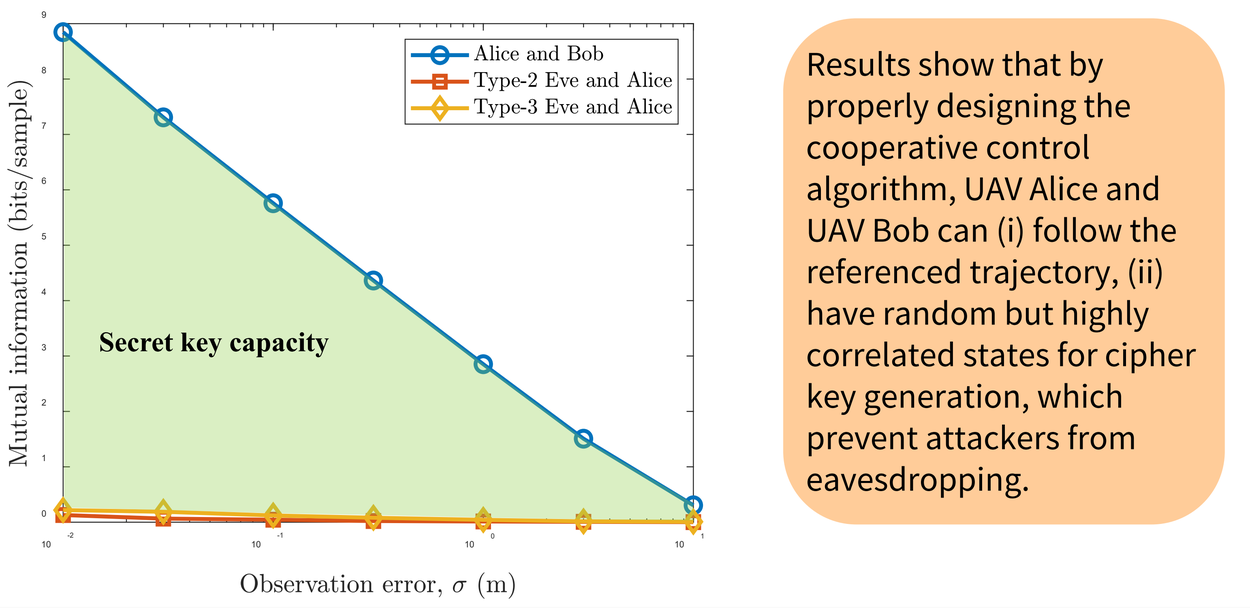
3. Implementation of Control Layer Security

Schematic Sketch



Simulation Results





Results show that by properly designing the cooperative control algorithm, UAV Alice and UAV Bob can (i) follow the referenced trajectory, (ii) have random but highly correlated states for cipher key generation, which prevent attackers from eavesdropping.

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Adversarial Attacks in image data and sensing: A threat to Autonomous Systems

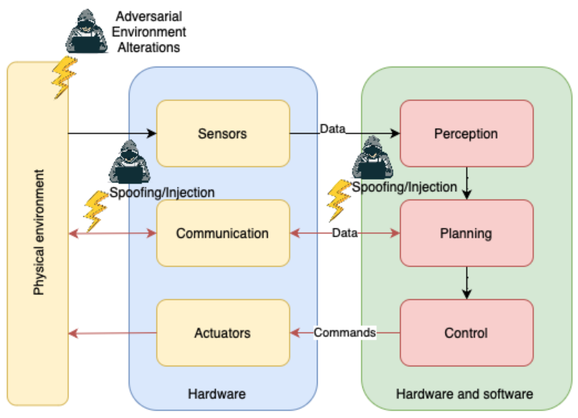
Lancaster University

Researcher: Alvaro Lopez Pellicer
Supervisors: Prof. Plamen Angelov, Prof. Neeraj Suri

Engineering and Physical Sciences Research Council

TAS-S

Attack Surfaces on Autonomous Systems Stack



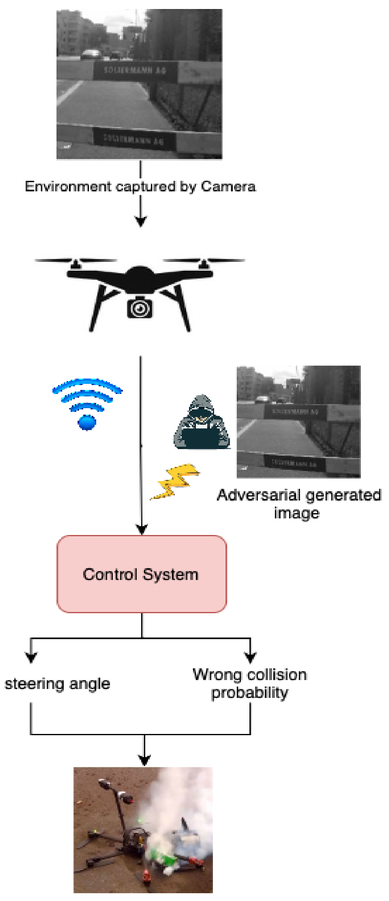
Autonomous systems face numerous challenges in their operation, due to the uncertain and dynamic multi-layer attack surfaces

Adversarial attacks undermine the security and trustworthiness of AS

These attacks can take various forms, such as data poisoning, model inversion, or evasion, and can have serious consequences for the safety, reliability, and privacy

Critical Impacts

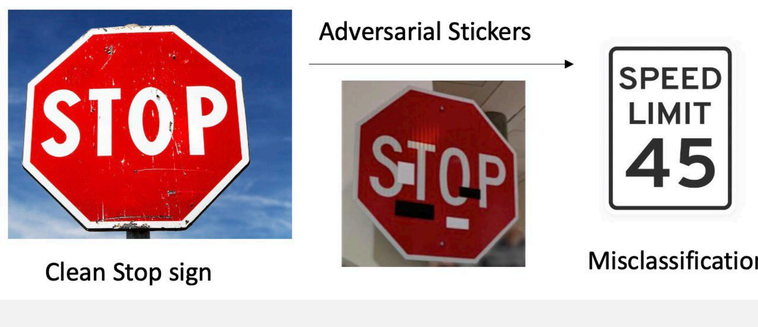
- **Perception layer**: Adversarial attacks can manipulate the sensory input of an AS, causing the system to perceive incorrect or misleading information. For example, adversarial examples in computer vision can cause an AS to misclassify objects in the environment, leading to incorrect or unsafe actions.
- **Planning layer**: Adversarial attacks can also manipulate the AS's decision-making processes, leading to incorrect or suboptimal plans. For example, an attacker may introduce false information about the environment or other agents, leading to incorrect or unsafe plans.
- **Control layer**: Adversarial attacks can also affect the control layer of an AS, leading to incorrect or harmful actions. For example, an attacker may manipulate the control signals or inputs to the actuators, causing the AS to take actions that are not in line with its intended behaviour.



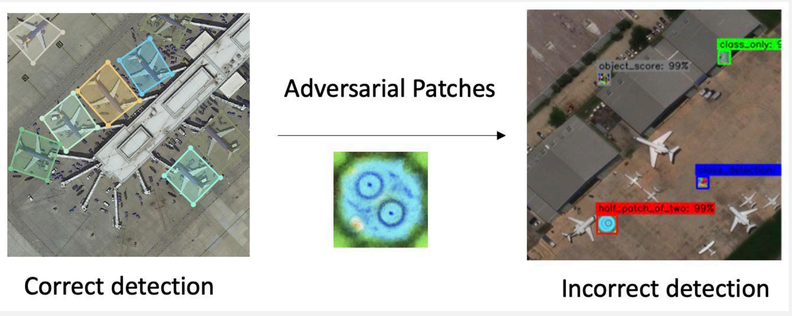
Attacks in the physical environment

Examples include:

Adversarial Stickers
✓ Target object classification



Adversarial patches
✓ Target object detection



Manipulate the environment in order to cause the system to behave in unintended or harmful ways.

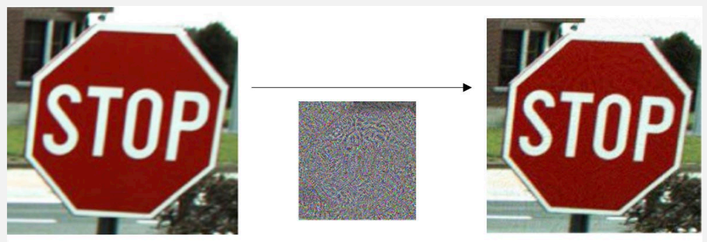
Attacks may be unnoticeable to humans when placed in the real world as they may be mistaken by decorations, urban art or vandalism and not seen as a bigger threat

Attacks on Digital Images

Example:

FGSM

$$\hat{x} = x - \epsilon \cdot sign(\nabla_x J(\theta, x, y))$$

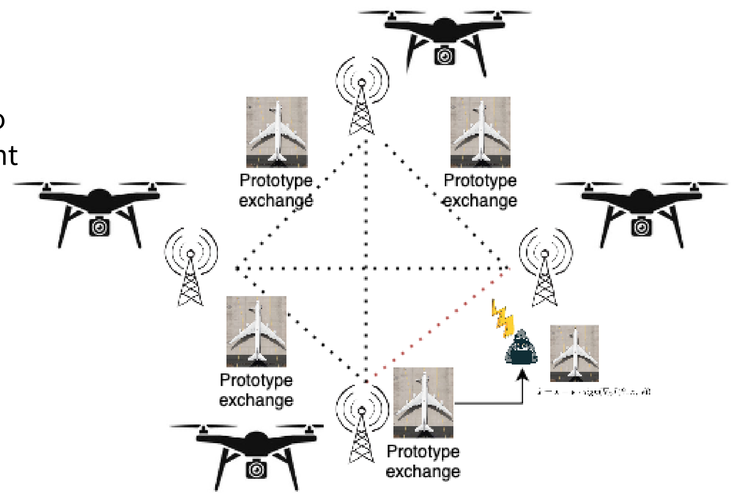


Creating small, carefully crafted perturbations to the input data in order to cause the machine learning model to produce incorrect or undesirable outputs.

Malicious perturbations in protype exchange in a FL environment

An attacker may use digital attacks to inject adversarial examples at different levels of a system such as in a distributed (Federated Learning) environment.

Given a Prototype based FL environment, threats may exist of spoofing in the prototype exchange stage with malicious images



Defence mechanisms

Different defence methods are being developed to tackle these challenges

They may be categorised as:

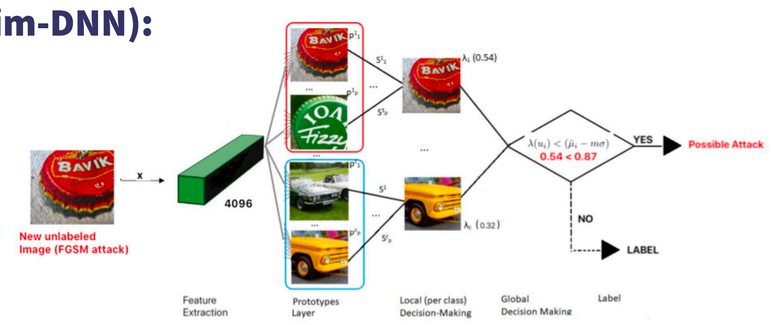
- Adversarial training
- Input data pre-processing
- Detection
- Provable
- Model ensemble
- Model distillation
- Hybrid defences

Requirements for robust to adversarial attacks systems in the context of AS:

- Able to detect attacks
- Able to react to detected attacks
- Evolve with new unknown types of attacks and situations

Similarity-based Deep Neural Network to Detect Imperceptible Adversarial Attacks (Sim-DNN):

Detect adversarial attacks through its inner defence mechanism that considers the degree of similarity between new data samples and autonomously chosen prototypes.



Future work

Robust to adversarial attacks evolving classification

- A prototype based framework able to detect and mitigate digital (noise based) adversarial attacks and learn from new classes
- Following the principle from Sim-DNN, this framework would be able to detect possible attacks or unseen classes.
- After detection, the flagged image will be inputted into a denoising framework which will remove adversarial perturbations (if any) and be able to determine whether the image was attacked or if it is a new unseen class and then create a new prototype for it

Advantages of the proposed framework

- Detect adversarial attacks with more confidence
- Mitigate detected adversarial attacks by removing the attack from the input and correctly reclassifying the image
- Evolving learning of new unseen classes

Disadvantages of the framework

- Potentially ineffective against physical attacks
- Will still have some of the drawbacks from Sim-DNN

Lancaster University

Cranfield University

UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Ensuring safe state space exploration of a Markov decision process (MDP) using Bayesian non-parametric (BNP) models for reinforcement learning (RL)

Lancaster University

Researcher: Xavier Hickman
Supervisors: Prof. Dan Prince, Prof. Neeraj Suri

Brief Overview of RL and Safe RL

Reinforcement learning addresses optimisation problems such as optimal control where other machine learning paradigms such as supervised and unsupervised fail [1]. Fundamentally optimal control is a field of mathematics that studies the problem of finding the best control strategy for a system, given a set of constraints and a criterion that defines optimality. Reinforcement learning algorithms have been shown to handle highly complex and uncertain environment dynamics which makes them well suited to a plethora of real-world applications such as autonomous vehicles and robotics. RL is also highly data efficient, in that it can learn from a limited quantity of data, which in many problem spaces is real barrier to entry for conventional ML paradigms.

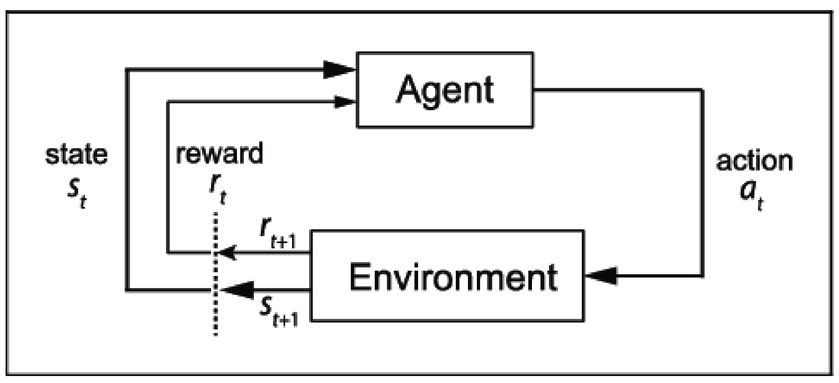


Figure 1: Schematic of Reinforcement learning Paradigm [1].

Safe reinforcement learning is a subfield of RL that focuses on the safety and reliability of reinforcement learning algorithms. The key motivation behind the discipline is the exploration-exploitation mechanism that enables conventional RL algorithms so learn. When training begins there is more weight given to choosing random actions to expand the known state space and nearer the end of training there is more weight given to choosing actions based existing knowledge of state space. This mechanism creates a problem in safety critical environments because the nature of the stochastic policy could potentially cause damage to the agent or the environment which in application contexts such as self-driving cars or military drones would be detrimental.

Safe RL Problem Formulation

Markovian sequential decision-making problems are often formulated as Markov decision process (MDP). MDPs are used across a variety of fields including RL, operations research and control theory. An extension of the MDP is the constrained MDP (CMDP) where the tuple includes a set of constraints which can be used to model properties such as safety. A formulation of a constrained Markov decision process is shown in figure 2.

$$M = \langle S, A, P, r(.,.,.), \gamma, C \rangle$$

Figure 2: Constrained Markov decision process [2].

Figure 3 shows is an alternative formulation of the safe RL problem. We aim to maximize the value function for some policy π of some state *s* at a given timestep *t* subject to the safety function evaluation of that state *s* and time step *t* being at or above some scalar threshold *h* which is problem specific [2].

$$maximise: V_N^{\pi}(s_t) = E\left[\sum_{t=1}^{N} \gamma^{t-1} r(s_t, a_t))\right]$$

$$subject\ to: g(s_t) \geq h, \forall t = [1, N].$$

Figure 3: Safe RL problem formulation [2].

Engineering and Physical Sciences Research Council

TAS-S

Why Bayesian non-parametric models?

Deep neural networks (DNNs) have been a very popular choice for function approximators in classical reinforcement learning in recent years [3][4], however Bayesian non-parametric (BNP) models offer some unique advantages over DNN's when optimizing for safety.

- DNN's have been shown to be very sensitive to distributional shifts in input data which in reinforcement learning problems is very common. Distributional shifts in observation data can occur for several reasons, but the most common is the non-stationarity of environments where the underlying distribution of states and actions change over time. In contrast BNP models are designed to be robust to distributional shifts in data and can learn flexible distributions that capture the underlying structure of the data.
- BNP models can capture and quantify uncertainty which is especially useful in safe RL as many environments where safe RL algorithms are applicable often have high levels of uncertainty. BNP models can use this when making decisions in these uncertain environments to ensure safe actions are taken whereas DNNs have no natural way of modelling uncertainty.
- DNNs are notoriously uninterpretable [5] whereas BNP models provide highly interpretable models as a result of their simpler model structure and transparent uncertainty approximations. The inherent interoperability can help in explaining an agent's behavior and actions which is especially useful in safety critical application contexts.

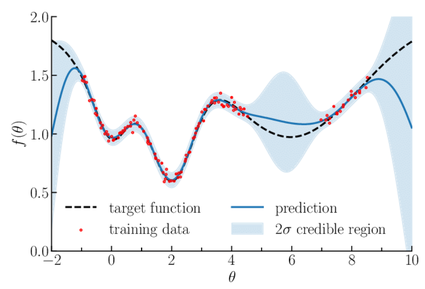


Figure 4: Illustration of gaussian process regression in 1 dimension [6]

Figure 4 shows an illustration of a 1-dimensional gaussian process which is a popular BNP model. The uncertainty is captured as the blue shaded area, so the larger the shaded are at certain points in the function the less certainty there is in the approximation.

Ongoing work

Stability of MARL in the face of perturbed communication:
Some of our ongoing work is looking at the stability and resilience of certain multi-agent reinforcement learning (MARL) algorithms in the face of severe network interruptions. These interruptions could affect the availability of communications mediums, or the integrity of messages sent. We are specifically interested in the multi-agent deep deterministic policy gradient algorithm (MADDPG) and the multi-agent proximal policy optimization (MAPPO) algorithm as they both demonstrate good performance on the multi-particle environments that best simulate groups of cooperating and competing autonomous systems [7][8].

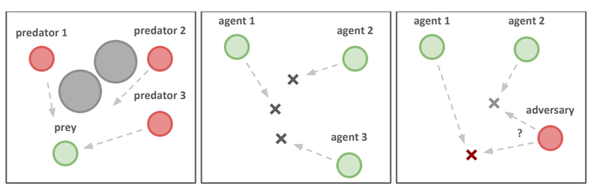


Figure 5: Screenshot of environments in the multi-particle environment (MPE) from the MADDPG paper [7]

Improving the SafeMDP algorithm:
Our ongoing work also includes a paper that is looking to improve the robustness of the SafeMDP algorithm [9] which utilizes a gaussian process to approximate safety in highly uncertain environments. SafeMDP provides desirable theoretical guarantees and demonstrates good empirical performance. We aim to use the SafeMDP algorithm as a base and develop a new algorithm which is based on a similar model to the gaussian process but with improved empirical performance w.r.t outlier observations and to build on the existing theoretical guarantees.

References

[1] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). The MIT Press.
[2] Akifumi Wachi, Yanan Sui, Yisong Yue, and Masahiro Ono. Safe exploration and optimization of constrained mdps using gaussian processes. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[3] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. , .
[4] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. (2017). Proximal Policy Optimization Algorithms.. *CoRR*, abs/1707.06347.
[5] S. Chakraborty *et al.*, "Interpretability of deep learning models: A survey of results," *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*
[6] Leclercq, Florent. (2018). Bayesian optimization for likelihood-free cosmological inference. Physical Review D. 98. 10.1103/PhysRevD.98.063511.
[7] Lowe, Ryan & Wu, Yi & Tamar, Aviv & Harb, Jean & Abbeel, Pieter & Mordatch, Igor. (2017). Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments.
[8] Yu, Chao & Velu, Akash & Vinitsky, Eugene & Wang, Yu & Bayen, Alexandre & Wu, Yi. (2021). The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games.
[9] Turchetta, Matteo & Berkenkamp, Felix & Krause, Andreas. (2016). Safe Exploration in Finite Markov Decision Processes with Gaussian Processes.

Lancaster University

Cranfield University

UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Research Activities: RS3A

RS3-Theme A: Behaviour Adaptation as a Basis of Security by Design

Lead: L. Dorn. Participants: J. Deville, A. af Wåhlberg.

Overview

As AS design and application progress, how will people adapt their behaviour in relation to them? And how might behavioural adaptations weaken AS security? Little is known about how critical aspects of a security breach may go unnoticed when operators are out of the loop.

To start with, we are focusing on autonomous vehicles and identifying security issues that may apply to other AS. Previous studies to evaluate behavioural adaptation in responses to assisted and automated vehicles have shown how unintended consequences can mean that safety benefits are not realised and may even be put at risk. These studies have been short term in duration and lab-based with very few studies conducted in real-world fully autonomous vehicles. Longitudinal field-based studies across a range of platforms, with younger and older people, will investigate behavioural adaptation to inform interface design in order to capture the operators attention when the security of the system is compromised.

Research activities

RS3A has been focusing on autonomous vehicles and identifying security issues that may apply to other AS. Our research activities have included:

- Rapid evidence review of methods used in research on autonomous vehicles and human behaviour adaptation. The main result of this was that the most common research methods are simulators and self-reports, while on-road studies are less than ten percent of the total. Studies are usually of very short duration, using small samples. This means that very little is known about how users of vehicles with autonomous features adapt their behaviour in the long term, and what this might mean for safety and security.
- Development of hypotheses about how trust in autonomous systems is built and lost, and a questionnaire to test these using validated scales. Distribution of the survey has begun with the aim of testing several different ideas about how trust in autonomous systems develops. The study will also investigate some methodological questions concerning this type of research methodology for evaluating trust.
- Building on the rapid evidence review, a comprehensive review of research on human interaction with autonomous systems is being compiled. Given the volume of the research in this field and building on the work of the rapid evidence review, this part of the research programme aims to divide the literature into three distinct areas: methodology, theory and empirical results pertaining to autonomous vehicles. These

Research Activities: RS3A

Research cont.

will be published as three journal papers, with submission during 2023.

- Investigation of various possibilities for accessing data and vehicles with autonomous features. Contacts have been taken with the AutoDrive project, Milton Keynes council, companies Aurrigo, Fetch and StageCoach operating an autonomous bus in Scotland.
- Involvement in the driver behaviour experiments of students on an MSc course at Cranfield. As these experiments require a human factors evaluation of the ADAS features of a Tesla, it is possible to gather data for the TAS-S project simultaneously with the students' work. However, to facilitate this, we organised participants, tasks, and measurement equipment. The study can therefore be considered a pilot, where problems with the reliability of the car, the simultaneous gathering of data for different purposes and technical issues with the measurement equipment were all prominent. These problems should be possible to solve and an on-road study on behavioural adaptation to ADAS functions be run during 2023. As this field study would include drivers with differing experience in ADAS, it will be able to quantify certain types of possible behavioural adaptation to ADAS. This study will thus approximate a field operational test of behavioural adaptation to ADAS.

Looking ahead...

RS3A has the following research activities planned for the next six to twelve months:

- Further work on the comprehensive literature review with a focus on automated vehicles given that most research has been conducted in this area and can inform us on the potential ways that humans may adapt their behaviour in response to an autonomous vehicle.
- A meta-analytic study on the correctness of safety forecasts for autonomous systems. There are many forecasts concerning how many accidents can be prevented by various types of technology, such as an Automated Braking System. These estimates are often given for several different ADAS systems and very large and when added together suggests we should not be experiencing any road traffic accidents at all. Yet, despite several different automated technological systems which aim to reduce collisions we still have many road traffic accidents, something would seem to be amiss. The forecasts may use erroneous assumptions, such as expecting that drivers will not change their ehaviour in response to these systems, For example, drive too fast in the belief that the ABS will protect them. Therefore, the values given are likely to be erroneous and can be tested in a meta-analysis allowing us to understand whether drivers appear to be adapting their behaviour in response to in-vehicle systems.
- To investigate this, forecasts of safety benefits for various technologies will be gathered and compared to results for these technologies. This will yield an indication of how correct current estimates of the safety potential of new automated features, such as Tesla's AutoPilot, might be. Furthermore, the discrepancy between forecasts and outcomes would be an indirect indicator of long-term behavioural adaptation, especially if

other types of crashes than those targeted have increased.

- Analysis of survey results. This includes effects of anthropomorphism, experience, and the general disposition of trust in technology. This work also includes cooperation with the Cranfield human factors in robotics team on previously gathered data.
- Investigation of autonomous shuttles in Sweden. There are two sites which run small, slow autonomous vehicles for public transportation (but still including safety drivers). Initial contacts with the researchers involved from Swedish universities indicate possibilities for cooperation.
- Contacts with StageCoach concerning their self-driving buses will be resumed in January 2023, when they hope to start operations in public transportation around Edinburgh. The investigations will mainly concern behavioural adaptation by the safety drivers and acceptance of the vehicles by the public.
- Field study on behavioural adaptation to ADAS features in vehicles with Cranfield University students on their MSc project.

Highlights & selected publications

Dr. Lisa Dorn was one of the experts working on the Technical Committee (TC 241) to develop a new Standard on guidance on safety ethical considerations for autonomous vehicles from July 2019 to September 2022. This committee sits under the BSI Road Traffic Safety management systems of the International Organization for Standardization. This forthcoming ethical guidance standard is called ISO39003 and approval is expected by the end of 2023**.**

Research Activities: RS3B&C

RS3-Theme B: Organisational Socio-Technical Mitigation

Lead: J. Deville. Participants: L. Dorn, C. May-Chahal, L. Moffat.

RS3-Theme C: Ethics and Governance of AS Security

Lead: C. Easton. Participants: L. Dorn, L. Moffat.

Overview

As public-private collaborations become more prevalent, there is a need to clarify the liabilities and duties of private companies working in a public capacity because there are different legal and incentive frameworks between private and public organisations. While these collaborations open new possibilities, they also bring forth a range of ethical, legal and social issues which warrant careful consideration. In a collaborative information management setting, it is important to support and encourage reflection on such issues by making more visible the ethical and legal implications of outsourcing, subcontracting, and privatisation in general.

As the rate of technological innovation exponentially increases, the ways that organisations manage their data, business, and their ethics, must adapt. This is not simply a case of 'keeping up' with the technology, but of creating synergies, affordances, and spaces for response. Given the extent and diversity of contexts in which A/S do and will operate, organisation adaptation needs to happen 'all the way through', from policy and protocol, to everyday practice.

Research activities

RS3B and RS3C have been collaborating closely to explore the challenges organisations have when designing and deploying, or responding to, Autonomous Systems (AS), with a focus on questions of security. Our partnership with National Highways (NH), has been particularly productive, with emergent insights on the challenges organisations confront when managing 'cultures of adoption' with organisational contexts, how organisations negotiate between different ethical frameworks in engaging with AS, as well as practical politics within organisations of engage with inherently uncertain futures as pertaining to AS vs. the pressure to implement solutions in the present.

A further area of focus has been on exploring how issues of AS security in relation to CAVs in particular are understood by the public. This has included RS3B and RS3C collaboratively designing a survey, with data collection led by IPSOS. This resulted in over 400 responses from members of the public. Emergent insights include a high degree of concern amongst the public about CAV safety.

Participants also raised concerns about the security of information flows as related to

Research Activities: RS3B&C

Research activities cont.

CAVs, with worries that AS technologies and their associated computer systems could become targets for hackers. Respondents also raised wider concerns regarding NH's role as a potential key player in taking responsibility for AS security on the UK's roads.

Looking ahead...

RS3B and RS3C have the following research activities planned for the next six to twelve months:

- Present emergent findings to RS1 & RS2 for implications for their work.
- Final phase research with NH, including 4 online focus groups with members of the public, drawn from those that responded to our earlier survey. Focused on adding richer, qualitative data to the survey results, and examining with participants specific 'participatory backcasting' scenarios involving AS deployment on UK roads.
- Final transition report, to be delivered to NH in spring 2023
- Confirm second project partner. This research would have two distinct areas of focus: (1) exploring the challenges posed to policing and investigative work around the increasing use of autonomous systems within vehicles, including but not limited to CAVs and (2) scoping the wider potential set of challenges that are raised for UK policing by AS as related to other technologies – drones, for example, whether used by members of the public, organised crime networks, or in commercial surveillance contexts.
- Organise 'Securing Trust in Autonomous Systems' workshop, in partnership with the Department of Organisation, Work and Technology at Lancaster University.
- We aim to submit two journal articles in Spring '23, 'Relational Approaches to Autonomous System Ethics', and 'From Makers to Publics: Bridging gaps between AS design and public perceptions'. The first examines relational and feminist ethics, focussing on the notion of entangled autonomy, that both humans and machines are only autonomous to the extent they participate in relations between others, environments, and things. The second concentrates on recent data generated from our partnership with National Highways, presenting some reflections on expert interviews and public focus groups about ensuring ethical and secure futures for autonomous vehicle use in the UK

Research Activities: RS3B&C

Highlights & selected publications

Collaboration with National Highways (please see pages 14-16).

- July 2022 ‘Relational Critiques of Autonomous Systems’, EASST ’22, Madrid
- Escalante, M.A.L, Moffat, L., and Büscher M. ‘Ethics through Design’, DRS2022, Bilbao, June 2022.
- Escalante, M.A.L., Moffatt, L., Harrison, L., and Kuh, V. (July 2021) ‘Dancing with the Troubles of AI’, Pivot, Online.

RS3: Unfolding the AV Dream

Lancaster University, Cranfield University

Researchers: Dr. Luke Moffat, Dr. Anders af Wåhlberg
Investigators: Prof. Corinne May-Chahal, Dr. Joe Deville, Dr. Lisa Dorn, Prof. Catherine Easton.

Engineering and Physical Sciences Research Council

TAS-S

Context

As part of the TAS Security Node, our work at Lancaster examines the User environment of Autonomous Systems (AS) particularly Autonomous Vehicles (AVs).
At Cranfield, we look at how and why the individual trust/use/accept autonomous technology like AVs, if the forecasts of safety of AVs can be trusted, and if behavioural adaptation (contra-productive actions) happens in AVs.

The AV dream

”Attacks”

Security measures can anticipate and/or prevent attacks

❖ Trustworthy

❖ Trusted

”Correct” use

The imaginaries of AV designers, manufacturers, and promoters, match those of users.
Users have correct mental models of the vehicles.
Driver behaviour match expectations of the makers

Safety & Security

Safety forecasts for automated features; crashes reduced by tens of percent

Roads and road users are safer

Convenience and Quickness

AVs increase efficiency on roads, they make people’s lives easier, more productive, and generate benefits for road networks

Description of map

The AV reality

Stockholm shuttle
Running since 2018
Six passenger seats
No seat for the safety driver
20 km/h max
Brakes hard for unknown reasons
Stops at every junction for OK from driver
10-15 passengers/day
Service a limited area like a taxi

Cranfield field study on behavioural adaptation using a Tesla postponed because the Autopilot feature of the car malfunctioned.

Do people trust AVs? Why not?

Tesla crash

❖ How is AV reality?

This vision from the fifties has still not come true

Have autonomous features of vehicles delivered expected safety gains, or has behavioural adaptation countered the effects?

Security & Ethics

These evolving realities open opportunties for rethinking technological imaginaries. Working with communities of potential users can help researchers understand, not just how they may or may not **adapt**, but what they **actually want and need** in their region.

Other ways of doing AS: Indigenous Protocols

One value from engagements with National Highways is Two-Spirit: encompassing bodily and intellectual relationships with AS technologies

Āina, from the Hawaiian for ‘land’

AI as Āina: ‘we should treat these relations as we would all that nourishes and supports us.’ (Lewis et al., 2018)

Lancaster University

Cranfield University

UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Autonomous Systems and Human Behaviour; Ongoing Studies

Cranfield University

Researcher: Dr. Anders af Wåhlberg
Investigator: Dr. Lisa Dorn

TAS Human Factors Work at Cranfield

- *Aim: Investigating how human interaction with autonomous systems can impact upon safety, especially when user behaviours change over time due to this interaction.*
- *Methods: Review, survey, meta-analysis, field study, (simulator).*
- *Status: Several studies under way.*

Interaction with AS: Review of Research

- Aim: Comprehensive review of research on how humans interact with AS in relation to safety. This includes three parts; methodology, theory and empirical results. This is the basis for all other work.
- Study 1: Methodology. To enable an understanding of what empirical results in the human-AS research area means, the methods which have been used to gather the data must be understood in terms of their validities and biases.
- Study 2: Theory. A bewildering array of different theories and concepts are used in AS-Human interaction research. These are summarized under some different headings and critically evaluated.
- Study 3: Empirical results. A traditional survey of available findings.
- Status: All studies under continuous but slow development, due to the literature being vast and other work taking precedence.

Multi-purpose survey

	Research questions	Method
1	Can anthropomorphism be primed and have an influence on survey responses?	Experimental manipulation by pictures of robots in two different versions of the survey
2	Can individual differences in trust be predicted by standard psychometric scales like Big Five personality?	Several different standard scales included, multiple regression analysis
3	Are there differences between active and passive acceptance of autonomous systems?	Items on passive acceptance (it's OK) and active acceptance (I will use) included. Comparisons of effect sizes of predictive scales.
4	Is intention to use AS predicted by Theory of Planned Behavior scales beyond what other scales included can predict?	Comparison of of correlations and partial correlations between intention and TPB predictors, the latter with other scales held constant.
5	Is trust in AS predicted by general trust, i.e. a general tendency to trust other entities?	Correlations between different scales.

Distribution and responses so far

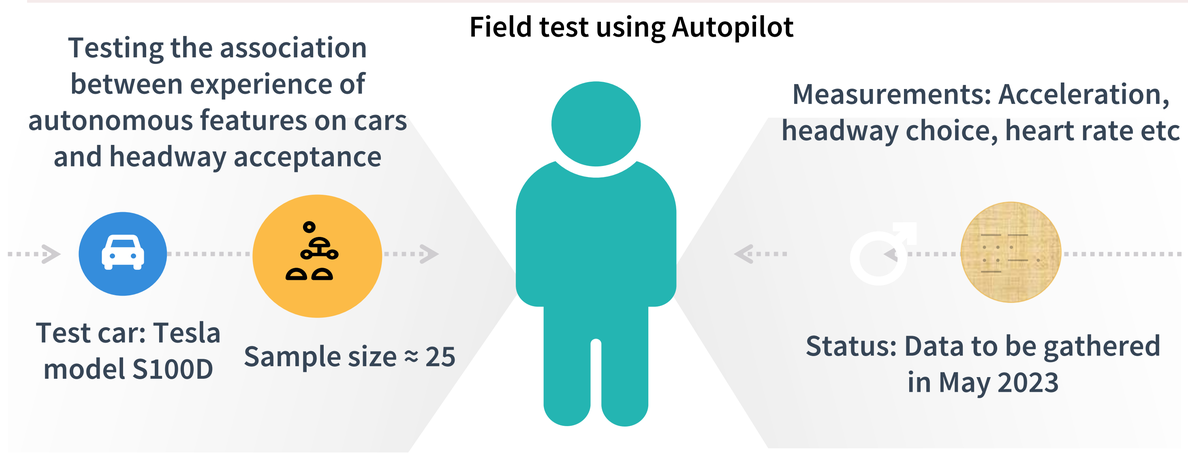
Groups targeted: TAS, Cranfield schools, Linkedin, ResearchGate, Facebook, Reddit.
Total N responses in February 2023: ≈150

Papers
Determinants of stated trust in autonomous systems (submitted)
Single factor or single source in self-reported trust in automation data? (in preparation)

Engineering and Physical Sciences Research Council

TAS-S

Field Study of Behavioural Adaptation to ACC



Meta-analysis of safety forecasts for vehicle automation

Safety is often predicted to increase by the use of automated features on vehicles – but how true are the forecasts claiming this?
"Vehicle stability control could prevent or mitigate up to 20 and 11 percent of moderate-to-serious injury crashes and fatal crashes, respectively." (Jermakian, 2012)
Such claims are based upon several assumptions, such as no change in driver behaviour due to the system implementation. However, behavioural adapation is always a possible threat to safety interventions, and have the potential to lessen the expected effect.

Studies: Effects for technological safety features on vehicles; forecasts versus reality in Australia (submitted)
Meta-analysis of forecasts and empirical investigations of technological safety devices for vehicles (in preparation)

Self-driving buses in the field; exploration

Autonomous bus to be launched in Edinburgh in 2023

Autonomous shuttles running in public transport in two places in Sweden

Possible explorative research questions:
What are the duties and behaviours of the safety drivers?
How have they been trained?
What are their beliefs about the systems they oversee?

Simulator study at Cranfield; under development

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Lancaster University
Cranfield University
UKRI Trustworthy Autonomous Systems Hub

AI Semiotics

Lancaster University

Researcher: Dr. Luke Moffat
Investigators: Prof. Corinne May-Chahal, Dr. Joe Deville

Engineering and Physical Sciences Research Council

TAS-S

AI Imaginaries & Realities

AI Semiotics & Semiotics of AI

As part of the TAS Security Node, my latest research has been exploring theoretical and methodological resources for confronting and understanding the new realities being created by AI proliferation. In particular, I investigate the ethical and social implications of AI, in how it is imagined, produced, used, and discarded.
Departing from AI semiotics, as a machine learning technique, semiotics of AI studies the cultural imagines associated with AI production
Below is a visualization of some of the fractal chains of exploitation required for AI production to happen. These chains are generative, meaning they create realities. And they are entangled, meaning when one changes, the others do too. These entangled relations are diffractive, non-linear, and foldable.

Credit: Serpinski; Phys.org, Crawford & Joler, 2018

Ethics through Design

Ethics through Design EtD is a cross-disciplinary, cross-sector framework for building a transition from ethical *technologies* to ethical *conduct*. Beyond 'by design' approaches, EtD focusses on the affective entanglements between people, ideas, and things, to argue for ethical conduct in all stages of A/IS design processes. Ethical value is not something that can be implanted into a device. In addition to specifications for trustworthy A/IS, EtD looks to the social dimensions of technology-use, and folds this together with calls for radical emancipation, and other ways of knowing, including indigenous protocols, traditional ecological knowledge, and feminist techno-science.

Less ethics, more politics

Scaffolding Dissent

Emphasizing ethical conduct brings things back from the heady space of techno-futurism, to the level of people and communities. How can communities have a say in the ways technologies are deployed where they live?

Some requirements:

- Empowering civic dissent
- Remembering the right to the city
- Considering tech and ecology as rooted in a common nature
- Amplifying marginalized and erased voices

Lancaster University

Cranfield University

UKRI Trustworthy Autonomous Systems Hub

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1]

Acknowledgements

We would like to thank our stakeholders for their continued support.

Engineering and Physical Sciences Research Council

This work is supported by the Engineering and Physical Sciences Research Council [grant number: EP/V026763/1].

Contact

TAS-S Node
Security Lancaster
InfoLab 21
Lancaster University
LA1 4WA

https://tas-security.lancs.ac.uk/
tas-s@lancaster.ac.uk
@TAS_Security