Journal of the International Society for the Systems Sciences | 68th Meeting of the International Society for the Systems Sciences
Wilby, J., Makar, J. and Friend, M. (eds)

AIC-BASED UNFORESEEN EMERGENT SCENARIOS (BLACK SWAN EVENTS) DISCOVERY METHOD FOR AUTONOMOUS SYSTEMS SAFETY DESIGN

Haider Al-Shareefy
ECS
University of Southampton
Southampton, UK
hal1e20@soton.ac.uk,
haideralshareefy@outlook.com

Michael Butler
ECS
University of Southampton
Southampton, UK
m.j.butler@soton.ac.uk

Thai Son Hoang
ECS
University of Southampton
Southampton, UK
t.s.hoang@soton.ac.uk

Abstract
This paper presents a new approach to enhancing the safety design of autonomous systems by applying systems science in identifying autonomous systems-related Unforeseen Emergent Scenarios (Black Swan events)—unexpected situations requiring critical decisions. Building on the AIC (Appreciation, Influence, Control) systems approach, this research incorporates a perspective shift to better analyse potential system failures in complex environments. The study introduces a systems thinking method with an AIC-based Chain-of-Thought (CoT) technique, which helps architects predict and analyse autonomous unforeseen behaviours that could lead to failures. By altering the type and direction of AIC interactions, architects can improve their predictive capabilities and prepare for complex interactions that traditional methods may overlook. The result is a list of Unforeseen Emergent Scenarios that can inform additional safety requirements, laying a strong foundation for designing safer autonomous systems.

Keywords
Systems Thinking, AIC Systems Approach, Chain of Thought, Safety Engineering, Autonomous systems

1 Introduction:

In order to ensure that autonomous systems designed to operate in open and complicated environments are trustworthy, architects require a deep understanding of the challenges posed by such complexity. These environments often involve the deployment of machine learning (ML)-based systems. For instance, an interesting incident was reported in Los Angeles where a food delivery robot was seen moving through a crime scene. (New York Post, 2022). The incident video shows a cameraman lifting the yellow tape to allow a robot to drive through a crime scene. In this interaction, an unintended relationship has formed between the robot and the cameraman, as their primary purposes have somehow cooperated, prompting the human to adopt an unexpected auxiliary purpose. As a result, the news reporters appreciated the robot's action, allowing the robot to move through the crime scene. This incident demonstrates a catastrophic failure; what if the robot tampered with or destroyed evidence that could solve the crime? We would classify this scenario as "Unforeseen" since such a situation is unforeseen by the architect's expectation. In prior research, we articulated a method to unpack the complexity of the autonomous systems problem domain (Al-Shareefy, Butler, & Hoang, 2024). The output of the latter process was a list of factors, assumptions and early identification of hazards. Then, we articulated a method to elicit system-level requirements based on the AIC systems approach (Al-Shareefy, Butler, & Hoang, 2023), and the output of that approach was a set of system-level AIC (Appreciation, Influence and Control) actions related to involved systems (not just the system of interest).

In this process, we systematically assist the architect in improving their capability to predict such hard-to-expect situations. We do so by building on the predicted AIC models from the requirements elicitation models and analysing them further to tease out potential unexpected behaviour. We include a modified AIC-based CoT technique (Al-Shareefy, Butler, & Hoang, 2024) that shifts the architect's perspective of any predicted relationship in the problem domain model. To demonstrate our approach, we used the Eagle Nest case study, which we used in our IS requirements elicitation process (Al-Shareefy, Butler, & Hoang, 2023). In particular, we took the output example of the latter paper and applied the current process to it.

The paper proceeds as follows: Section 2 outlines the theoretical foundations, Section 3 presents the methodology, and Section 4 discusses a case study and results. Section 5 summarises the conclusions and implications for future research.

2 Background:

The structured approach presented in this paper directly supports adherence to ISO 15288's Business and Mission Analysis process (INCOSE, 2023) by helping architects define the strategic problem space for autonomous system safety, characterising the solution domain, and identifying alternative solution classes that address emergent risks. Our approach relies on two main concepts:

2.1 AIC Systems Approach:

The AIC (Appreciation, Influence, Control) Systems Approach is a system science-based systems engineering approach that builds on the AIC systems theory (Smith W. E., 2016). We introduced the AIC systems engineering approach in our work (Al-Shareefy, Butler, & Hoang, 2023). It offers a framework for understanding and engineering complex systems' power dynamics in any environment. AIC theory emphasises that power (which we interpret as a system's capabilities or actions) is central to organisational dynamics in systems and derived from organisational purpose. Inspired by AIC assumptions, our systems approach presumes that any system behaves as if it is pursuing some Primary Purpose (PrimeP). In order to achieve its PrimeP, it must identify three types of auxiliary goals and powers (interpreted as actions) that any system in any environment ought to establish to ensure balance and longevity of existence in that environment. Our approach defines minimum requirements to form a relationship in some environments consisting of a source system (source of purpose), a receiving end sink system, a goal and an action. There are three types of goals and actions: Appreciation, Influence and Control. Actions ought to be effective in delivering the intended goal, and goals intend to deliver purpose. There are three types of emergent effects: Obstructive, Supportive and Neutral. Given the earlier articulation, we define three types of relationships:

- **Control relationship:** direct causation without persuasion or preceding actions. For instance, a drone perception and guidance system directly sets priorities for the drone flight controller's actions.

- **Influence relationship:** indirect causation with a higher degree of uncertainty in achieving such causation since it relies on controlling a third-party situation, which in turn controls or influences the target system. For example, the perception and guidance system indirectly controls (influences) the propulsion system (by controlling the flight controller, which controls the propulsion).

- **Appreciation relationship**: Appreciation means managing the unavoidable influence of factors or features beyond the control or influence of the appreciating system. For example, the drone flight controller must appreciate wind direction and speed to effectively control the flight propulsion system.

Systems' behaviours are understood by defining what they control, influencing what is outside direct control, and appreciating factors that can neither be influenced, controlled, nor avoided. Understanding these dynamics helps grasp the internal flow of power and how an organisation interacts with its environment.

2.2 Systems thinking-based CoT methods:

Another important aspect the reader needs to know is; systems thinking-driven CoT methods. The concept employs the prompt engineering CoT technique as an objective method to apply systems thinking and science where needed in intelligent-systems engineering design. In the realm of AI and NLP, CoT techniques have been pivotal in enhancing the capabilities of algorithms, especially in tasks involving deep reasoning and complex scenario analysis. Research indicates that explicating the reasoning process step-by-step significantly augments the interpretability of AI systems, making them more reliable and easier to debug and validate (Chu, et al., 2023) (Wei & al, 2022). Each step in the systems thinking-based CoT process involves formulating a predictive question, based on implementing some general systems rule, followed by a guiding prompt that directs the analysis, helping unravel the intricacies of the system's operation and environment. Systems Thinking CoT attempts to provide a construct that applies general systems theories or assumptions to enable architects to formulate systems-objective situational awareness. An inspiration was taken from the "game of life" model whereby a set of interrelated general rules can objectively produce a variety of complex variations from a given simple structure.

In our previous work (Al-Shareefy, Butler, & Hoang, 2024), we integrated the AIC framework into the CoT process, representing a method in applied systems thinking. Known as the AIC-based CoT, this method applies the principles of the AIC systems approach to bolster confidence in predictive reasoning about system behaviours and outcomes. The process is iterative, where each step builds upon the previous ones, refining assumptions and deepening the understanding of the system's dynamics. The AIC CoT systems approach starts with taking input from some system interaction descriptions. The first step prompts the architect to identify the list of unsafe behaviours. Then, it prompts the architect to question what systems are involved. After that, the approach prompts the architect to question the PrimeP of the observed system and follow up by defining how the system achieves the primary purpose via auxiliary AIC relationships. To define a systems thinking-based CoT in general, the architect needs to define a CoT title, CoT's primary purpose and CoT's input. Also, the general systems perspective (rule) needed to be applied to make a prediction. Furthermore, the CoT structure included the thinking process and the expected output.

3 Materials & Methods:

In this paper, we focus on discovering Unforeseen Emergent Scenarios. An Unforeseen Emergent Scenario, which also can be referred to as "Black Swan Event" (Kalia, Menzel, Grello, & Walker, 2024), is a less expected situation that may occur in the real world where an autonomous system is expected to make the right decisions—for example, the food delivery robot in a situation like the one described above. The term "Edge Cases" (Karunakaran, Worrall, & Nebot, 2020) is commonly used to describe hard-to-expect-scenarios. However, we believe such characterisation is an inappropriate adaptation of a concept well understood in deterministic software development where maximum and minimum input boundaries are clearly identifiable. For non-deterministic systems such as ML-based perception, there is no clearly identifiable boundary of inputs for pictorial datasets. For example, how can we specify a "human" in a pictorial sense to clearly define maximum or minimum edge cases? Using the term "Unforeseen Emergent Scenarios" is more appropriate, for example, when facial recognition datasets did not consider people wearing COVID-19 protective masks.

3.1 Research Problem:

Predicting Unforeseen Emergent Scenarios is crucial for robust design and operational resilience in complex systems architecture. Common hazards and risk analysis techniques may help developers to discover unsafe scenarios. However, more work needs to be done regarding autonomous systems design since the uncertainty of emergence is exponentially more unpredictable than non-autonomous systems. The AIC (Appreciation, Influence, Control) perspective shift technique, as applied within the AIC General Systems Approach (GSApp), enables system architects to predict potential hard-to-expect unsafe scenarios by shifting perspectives within known events. This method triggers the architect's predictive thinking by reimagining events from an alternative worldview within the AIC framework.

3.2 Identifying Unforeseen Emergent Scenarios Methodology (Proposed Solution):

Our method integrates AIC-based predictive modelling with CoT analysis to enhance safety design:

Given Model of the system interactions using the AIC modelling schema as an input (Al-Shareefy, Butler, & Hoang, 2024). Then, the following are the main activities for identifying Unforeseen Emergent Scenarios:

Step A) Perform AIC factorisation of the systems phenomenon described in the requirement.

Step B) Apply an AIC perspective shift to predict Unforeseen Emergent Scenarios. Four types of perspective shifts can be performed (see sections 3.3 and 3.4).

The output can be used to update system-level safety requirements to account for newly discovered risks. This approach enables early-stage identification of risks that could lead to system failures in unpredictable environments. Below is a detailed description of the AIC perspective shift methods.

3.3 AIC perspective shifts methods:

The AIC perspective shift leverages a multi-dimensional approach within the AIC framework, facilitating a reevaluation of known events or interactions from alternative perspectives. This technique comprises four types of shifts: altering the AIC type of an interaction, reversing the direction of the interaction, altering the type of effect for an action and altering all of the types together.

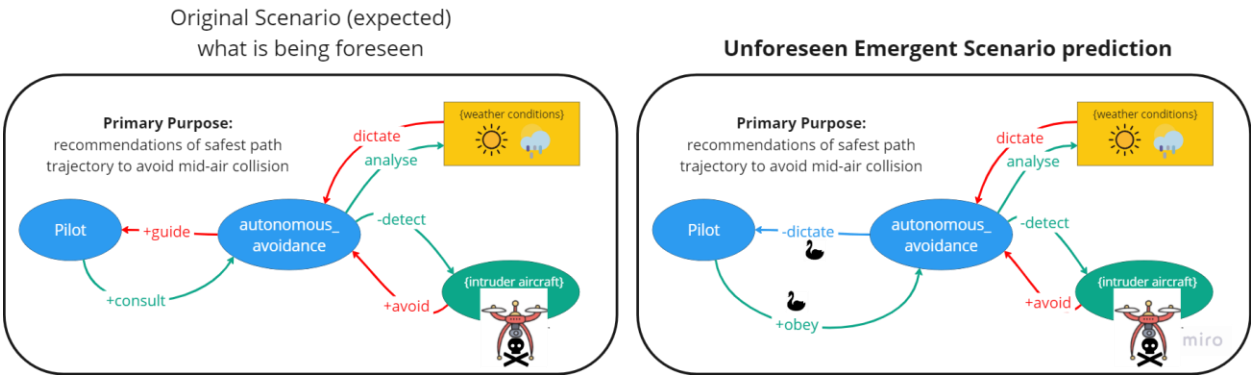
3.3.1 Predictive Perspective Shift by Altering Interaction's AIC Type:

This dimension entails modifying the AIC type of interaction within a system, transitioning an element's role from appreciation to influence or control or vice versa. This allows architects to explore how changes like interaction affect system behaviour and outcomes. The shift involves redefining the functional goal of an element in a scenario, for example, changing an advisory tool into an authoritative controller. This transformation prompts a re-evaluation of system dynamics and user interaction patterns.

Example;

Consider an aircraft avoidance recommendation system (source) initially designed to influence the pilot's decision-making process (sink) in a supportive manner. Applying the perspective shift, the system's interaction is reimagined as controlling, potentially obstructing the effectiveness of the pilot's intuition role in the autonomous decision-making process. Exhibit 1 graphically depicts the original scenario of interactions and the predicted potential scenario of interactions over time.

Exhibit 1 AIC modelling schema demonstrating shifting an established perspective (left side). Interaction lines: Red means influence, blue means control, and green means appreciation.



Unforeseen Emergent Scenario Prediction;

- **Initial scenario (Influence)**: The relationship between the recommendation system and the pilot is primarily influential and supportive. Pilots might initially perceive the system as somewhat unreliable, preferring to rely on their intuition.

The perspective is then can be shifted to view this system as a controlling entity:

- **Shifted perspective scenario (Control)**: If we alter the perspective, we can envision the recommendation system as directly controlling the pilot's behaviours, particularly in high-stress situations where the pilot may depend entirely on the system's directives, potentially to the detriment of safe flying practices. Such a situation can be considered as a surprising, Unforeseen Emergent Scenario.

3.3.2 Predictive Perspective Shift by Reversing Interaction Direction:

This approach reverses the flow of interaction, transforming the source of an action into its recipient (sink), and vice versa. This reversal thought, challenges conventional flow dynamics within the system, prompting new insights into potential system behaviours and vulnerabilities. Reversing the interaction direction can reveal vulnerabilities or strengths in the system by showing how changes in the flow of information or control can lead to different outcomes.

Example;

Weather conditions (source) influence an aircraft collision avoidance recommendation system (sink). The system must appreciate the weather when attempting to control the types of recommendations it makes while influencing the pilot's decision-making process to avoid a potential collision.

Unforeseen Emergent Scenario prediction;

- **Initial scenario (Influence)**: An initial interaction between the weather conditions (appreciated sink) and an aircraft collision avoidance recommendation system (appreciating source), with the recommendation system being influenced by environmental conditions.

By reversing this interaction, the aircraft can be imagined as an influencer of the environment:

- **Shifted perspective scenario (reversed influence)**: If we alter the influence direction, we have a situation where the aircraft influences the environment. This would be counterintuitive and surprising; what kind of scenario could the aircraft be viewed as having influenced some aspect of the environment? This might involve the aircraft's systems modifying environmental data inputs due to a cyber attack or a bug, which could lead to erroneous readings of sensor data being fed into navigational or operational systems.

3.3.3 Predictive Perspective Shifts by altering types of effect [Supportive, Obstructive, Neutral]:

The perspective shift by altering the type of effect involves changing the impact that the actions of one system (source) have on another system (sink) in terms of being supportive, obstructive, or neutral. This shift allows system architects to evaluate the adaptability and resilience of systems under different operational conditions by envisioning how variations in the effect type could alter behaviours and outcomes. This analysis is crucial for understanding a system's potential vulnerabilities or strengths that may not be evident under normal conditions.

Example;

In its standard operational mode, an aircraft collision avoidance system is designed to enhance safety by providing timely warnings and manoeuvre suggestions to avoid potential aerial conflicts. The system continuously monitors the airspace for other aircraft and uses algorithms to predict potential collision paths. It then alerts pilots with specific avoidance manoeuvres.

Unforeseen Emergent Scenario prediction;

- **Initial scenario (supportive effect)**: The system supports the pilot by augmenting their situational awareness and decision-making capabilities, helping to maintain safe distances from other aircraft and navigate congested airspaces effectively.

The perspective is then can be shifted by re-imagining the type of effects, changing from supportive to obstructive or neutral:

- **Obstructive Effect**
    - **Shifted perspective**: By altering the perspective to an obstructive effect, we can reimagine the collision avoidance system as temporarily hindering the pilot's ability to execute certain manoeuvres.
    - **Operational Scenario**: During a high-density traffic situation at an airport, the system might excessively restrict pilot manoeuvre options to prevent potential collisions. This might force pilots to follow less optimal paths or delay landings, which could disrupt flight schedules and increase fuel consumption.
    - **Impact**: Here, the system, usually supportive, now acts as an obstruction to efficient flying, prioritising safety over efficiency and convenience.
- **Neutral Effect**
    - **Shifted perspective**: Shifting to a neutral effect, the system could be imagined as non-responsive or minimally interactive, neither aiding nor hindering the pilot actively.

- **Operational Scenario**: Consider a scenario where the system enters a diagnostic mode during flight, ceases to provide active collision warnings, but continues monitoring airspace without issuing commands.

- **Impact**: In this mode, the system neither supports nor obstructs the pilot, leaving them to rely entirely on their training and other navigational aids. This could be particularly challenging during low visibility conditions or in highly trafficked airspaces.

3.3.4 Predictive Multi-types Shifts by Altering AIC and effects types, as well as interaction direction:

This comprehensive shift combines altering the AIC type and effect type and reversing the direction of interaction, providing a robust method for reevaluating system interactions. This multidimensional shift enables the exploration of complex scenarios where changes in system operational dynamics might occur. By simultaneously shifting the AIC type and interaction direction, architects can simulate complex interactions that might not be apparent under normal analysis conditions.

Example

Initially, the aircraft avoidance recommendation system influences pilot decision-making:

- **Initial scenario**: The system provides recommendations, and the pilot uses these to make informed decisions.

The scenario is then reimagined with both the AIC type and interaction direction altered:

- **Shifted perspective scenario**: The pilot controls the recommendation system's sensitivity settings, directly manipulating its operational parameters and recommendations. This thought process may prompt us to question what kind of scenario in which the pilot side controls recommendation parameters. Does this mean we need to give the pilot such a level of control? Also, it may prompt us to question whether there may be a cyber attack that exploits pilot access to the recommendation system.

3.4 Predictive AIC perspective shifts CoT:

In order to apply the concepts above, we defined AIC perspective shifts CoT in Exhibit 2. The process outlined in Exhibit 2 involves a methodology aimed at predicting Unforeseen Emergent Scenarios from anticipated system interactions by shifting perspectives based on AIC factors. Such a technique is aimed at assisting architects in discovering unsafe interactions in hard-to-predict operations domains. Given a description of a phenomenon and assuming a pre-determined AIC factorisation of relationships, these relationships are expected to change due to evolving complexity. The process involves four steps: 1) assessing what happens if the interaction flow and effect type remain constant while altering the AIC type of purpose and action; 2) considering the impact if the interaction purpose and action type stay the same but the interaction flow direction reverses; 3) exploring the outcome if the interaction purpose type, action type, source action, and interaction flow direction remain constant but the action effect changes; and 4) evaluating the scenario when all AIC factors (interaction purpose, action type, source action, and interaction flow direction) change. Each step is guided by specific prompts to redefine actions and scenarios within the shifted context, ultimately generating four unforeseen scenarios for each system phenomenon described.

Exhibit 2. Predictive AIC perspective shifts CoT

CoT Title	AIC perspective shift
CoT Primary Purpose	To assist with predicting unexpected, Unforeseen Emergent Scenarios from anticipated systems interactions
CoT Input	A description of some phenomenon (an interaction or a scenario)
General Systems Rules	Given that universal disorder, thus variations increases over time, it is expected that a certain AIC-formulated relationship of some complex with its environment will change unexpectedly due to a change in complexity over time.
CoT Process	**Step 1)** What would happen if the interaction flow and effect type remained the same, but the AIC type of purpose and action type were altered in the future? **Guiding prompt:** Review the AIC dynamics of the observed systems and alter the nature of purpose and action. Then, define an appropriate action to bear the new AIC types and describe a scenario in the shifted context.
	Step 2) What would happen if the interaction purpose and action type remained the same but the interaction flow direction reversed in the future? **Guiding prompt:** Review the situation of the observed system and swap only the interaction direction, making the source the sink and the sink the source of the new potential interaction. Then, define an appropriate new action to bear the initial AIC types but with a reversed interaction flow direction and describe a scenario in the shifted context.
	Step 3) What would happen if the interaction purpose type, action type, source action, and interaction flow direction were the same but the action effect changes in nature in the future? **Guiding prompt:** Review the situation of the observed system and only change the effect type (supportive, obstructive, or neutral). Then, describe a scenario in which the nature of the source action is different.
	Step 4) What would happen if all AIC factors' interaction purpose, action type, source action, and interaction flow direction were changed in the future? **Guiding prompt:** Review the situation of the observed system and describe a scenario whereby the purpose and action types (appreciation, influence, control), as well as the effect type (supportive, obstructive, or neutral), are all changed. Reverse the direction of the interaction and the action itself too.
CoT Output	Four different unforeseen scenarios for every given system phenomenon.

4 Application to The Eagle Nest Case Study and Results:

In this section, we will directly formulate and apply the CoT method (presenting the application of only step 1 of the CoT in this work), which is driven by the AIC systems approach, to the Eagle Nest case study. Eagle Nest is an autonomous system comprised of an intelligent security drone and docking station to secure a train track zone. To achieve such PrimeP, the autonomous drone (Eagle Drone) is also designed to recognise and avoid adversarial drones using a machine learning-based perception system while conducting security patrols along the train tracks. The input to the CoT method is a system-level requirement. We will use an output of our IS systems requirements elicitation method. Our analyses discovered an Unforeseen Emergent Scenario where local people may use laser attacks to disorient the Eagle Drone while conducting its security patrol. To understand how the requirement was developed, review our work here (Al-Shareefy, Butler, & Hoang, 2023):

The Eagle Drone camera system shall automatically turn away from the direction of any detected laser attack and then return to its original position in order to maintain a secure patrolling zone along the boundary fence line.

4.1 Step A) Perform AIC factorisation of the systems phenomenon described in the requirement:

The first step is identifying the AIC interactions involved in the requirements. The purpose, in AIC approach, is the main aspect that is being served by the requirement and is always identified by the phrase "in order to …". To manifest the purpose, identify the list of possible single interactions that impact the achievement of the purpose. A single interaction must have only a source, a sink, and an action. Apply AIC factorisation CoT of the interaction. To define the list of interactions, consider that the requirement describes a phenomenon, then break down the phenomenon into a list of required atomic interactions which give rise to the phenomenon. For example, a drone operates propellers to take off. The latter is a phenomenon that involves a drone, propellers, and taking off. What list of interactions is required to give rise to the phenomenon? Mainly, the drone switches on the drone operating system, the drone activates the flight controller, the flight controller operates the motors, the motors rotate the propellers, and the propellers propel the drone upwards. Exhibit 3 captures the list of interactions and their AIC factorisation.

Exhibit 3. Performing AIC factorisation of system requirement

Interaction	**AIC analysis**	
Maintaining bounded train track zone patrolling capability avoids laser attack	**Source**	Camera perception
	Sink	Laser attack
	Action	Avoid
	Action type	Appreciative
	Purpose type	Appreciation
	Effect type	Obstructive
	Source	Maintaining bounded train track zone patrolling

Maintaining bounded train track zone patrolling capability adjusts the camera orientation	**Sink**	Camera orientation
	Action	Adjust
	Action type	Controlling
	Purpose type	Control
	Effect type	Supportive
Maintaining bounded train track zone patrolling capability steers flight direction	**Source**	Maintaining bounded train track zone patrolling
	Sink	Flight Direction
	Purpose type	Control
	Action type	Controlling
	Action	Steer
	Source action effect on sink	Supportive
The camera system perception detects laser attack light	**Source**	Camera perception
	Sink	Laser attack
	Purpose type	Appreciation
	Action type	Appreciative
	Action	Detect
	Source action effect on sink	Obstructive
The laser attacks the camera system	**Source**	Laser attack
	Sink	Eagle drone
	Purpose type	Influence
	Action type	Influencing
	Action	Disorient
	Source action effect on sink	Obstructive

4.2 Step B) Apply AIC perspective shift to predict Unforeseen Emergent Scenarios:

The 4 modes of AIC perspective shifts are applied for every identified interaction. To help imagine new scenarios, you may consider HAZOP keywords: less, more, part of, early, late, no or not, and reverse flow of information, command, or substance. We will use the AIC perspective shift analysis of the "the camera perception system detects laser attack light" interaction as input for applying AIC perspective shift CoT.

4.2.1 Step B.1) Perspective Shift by altering interaction's AIC Type:

In this method, we will fix the initial perspective of all AIC factors and alter the AIC purpose and action type from appreciation to control. Then, we will reimagine a situation where the perception system controls the laser attack's behaviour. We can think of "reflecting," considering that a reflective material could potentially coat the camera lens. This realisation then allows us to predict a potential Unforeseen Emergent Scenario as defined in Exhibit 4.

Exhibit 4. AIC-type perspective shift

Interaction	The camera system perception detects laser attack light		
AIC factors	**Initial perspective**	**Shifted Perspective**	**Unforeseen Emergent Scenario**
Source	Camera perception	Camera perception	The camera perception may behave as if it seeks to reflect the laser attack beam by orienting the camera lens (coated with a reflective coating) to reflect the laser attack away.
Sink	Laser attack	Laser attack	
Purpose type	***Appreciation***	***Control***	
Action type	***Appreciative***	***Controlling***	
Action	Detect	Reflect	
Source action effect on sink	Obstructive	Obstructive	

Here is a step-by-step explanation of how this perspective shift leads to the derivation of an unforeseen emergent scenario:

1. **Purpose Type: From Appreciation to Control**

    - **Initial Perspective**: Initially, the purpose of the camera system's interaction with the laser is appreciation. This means the system acknowledges the laser's presence but cannot influence or alter the laser attack impact or whatever is controlling it.

    - **Shifted Perspective**: The purpose shifts to control, indicating a proactive approach, a presence of capability, an emergent one where it is not expected, to manage the impact of the laser. Instead of simply detecting the laser, the system now aims to mitigate or eliminate its effect.

2. **Action Type: From Appreciative to Controlling**
    - **Initial Perspective**: The original action is appreciative, confined to detecting or recognising the laser attack.
    - **Shifted Perspective**: The action becomes controlling, involving direct intervention to alter the state of the laser attack, specifically through reflecting it.
3. **Action: From Detect to Reflect**
    - **Initial Perspective**: The fundamental action is to detect the laser light, a passive form of perception.
    - **Shifted Perspective**: This changes to "reflect," which is an active and direct method of manipulating the laser light to prevent it from achieving its intended effect.

5 Discussion & Conclusion:

This paper critically evaluates the effectiveness of the AIC-based Unforeseen Emergent Scenarios Discovery Method for autonomous systems safety design. The method has demonstrated a good potential in identifying hard-to-predict scenarios that conventional models might overlook. Integrating the AIC framework with the CoT technique provides a systematic and useful framework for dissecting and re-evaluating system interactions. The general systems thinking approach underpinning the AIC framework leverages generic intuitions about how systems behave, making it broadly applicable across diverse fields and complexities. This generic foundation is particularly advantageous in environments with high uncertainty and rapid technological changes. The adaptability of the approach is evident in its application to both traditional safety-critical systems and modern autonomous systems integrating AI and ML technologies.

5.1 Effectiveness in Discovering Hard-to-Predict Scenarios:

The method's application to the Eagle Nest case study demonstrates its effectiveness in uncovering scenarios that might not be immediately apparent. By applying AIC perspective shifts, the research team successfully identified multiple potential scenarios, such as the possibility of the camera perception system not only detecting but also reflecting laser attacks using a coated lens. This scenario, derived by altering the interaction's AIC type from appreciation to control, illustrates a proactive system adaptation in response to external threats. Further, the reversal of interaction direction in another AIC shift posited a scenario where the laser attack could potentially "detect" the camera system, suggesting the development of adversarial intelligence that adapts based on the defensive tactics of the drone. This scenario underscores interactions' dynamic and adversarial nature in intelligent security systems, where both the system and its threats evolve.

5.2 Key Contributions

- Introduced an AIC-based predictive CoT analysis framework. Leveraging CoT and predictive AIC systems thinking modelling.
- Demonstrated its application in autonomous drone surveillance.
- The output can be used to improve system safety through early risk identification of Black Swan events (for example, during the Mission and Business analysis process).

Our findings suggest that integrating systems science with AI safety engineering offers a promising direction for mitigating Black Swan events in autonomous operations.

5.3 Limitations and Future Research

While the AIC-based approach enhances predictive safety capabilities, it relies on accurate system modelling and predefined interaction parameters, which may not fully capture highly dynamic or adaptive autonomous behaviours. Additionally, the computational complexity of analyzing a large number of potential emergent scenarios may limit real-time applicability, necessitating further research into scalable and automated scenario generation techniques.

6 Related work

HAZOP (Hazard and Operability Study) (Crawley, 2015) is a structured and systematic technique for system examination and risk management that identifies potential hazards and operational issues using guidewords to investigate every conceivable deviation in system parameters. While effective in identifying deviations from intended operations, HAZOP requires very clear, direct, hard interactions among components. Guide words may not be meaningful to cover enough variety of other types of interactions.

STAMP (Systems-Theoretic Accident Model and Processes) (Leveson, 2011) focuses on systems control and constraints, analysing potential failures regarding inadequate enforcement of constraints and control flaws. While advanced in its approach to system safety through a control structure perspective, STAMP requires comprehensive systems modelling and may overlook specific user or environmental interactions that do not directly relate to control structures. On the other hand, The AIC framework's flexibility in adjusting interaction types and effects provides a broader lens for identifying unforeseen scenarios, including those outside of direct control loops, accommodating more varied system-user-environment interactions.

FRAM (Functional Resonance Analysis Method) (Hollnagel, 2012) models complex socio-technical systems by analysing the variability of everyday performances and how these can resonate to produce emergent outcomes. FRAM's focus on variability can make it challenging to apply in highly technical or less human-centred systems where variability is less about human performance and more about technological interactions. AIC's capability to systematically modify and shift interaction perspectives effectively encompasses human and technological variabilities, enhancing the predictive analysis of system behaviours.

Acknowledgements: Al-Shareefy is supported by a Thales EPSRC iCASE Award. Butler and Hoang are supported by the HD-Sec project, which was funded by the Digital Security by Design (DSbD) Programme delivered by UKRI to support the DSbD ecosystem.

References

Al-Shareefy, H., Butler, M., & Hoang, S. (2023). AIC Approach for Intelligent Systems Requirements Elicitation. *7th International Conference on System Reliability and Safety (ICSRS).* Bologna. doi:doi: 10.1109/ICSRS59833.2023.10381076

Al-Shareefy, H., Butler, M., & Hoang, S. (2024). An AIC-based approach for articulating unpredictable problems in open complex environments. *19th European Dependable Computing Conference.* Leuven, Belgium. Retrieved from https://arxiv.org/abs/2403.14697

Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., & al., e. (2023). *A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future.* Cornell, Available: https://arxiv.org/abs/2309.15402: Cornell University arXiv.org.

Crawley, F. (2015). *HAZOP: Guide to Best Practice.* Glasgow, Scotland: Elsevier.

Hollnagel, E. (2012). *FRAM: The Functional Resonance Analysis Method.* London: CRC Press.

INCOSE. (2023). *Systems Engineering Handbook* . INCOSE, ISBN: 978-1-119-81431-3.

Kalia, P., Menzel, M., Grello, K., & Walker, A. (2024). Integrated Systems Engineering, Safety, Reliability and Risk Management – Minimizing Black Swan Events. *2024 Annual Reliability and Maintainability Symposium (RAMS)* (pp. 1-6). Albuquerque, NM, USA: IEEE.

Karunakaran, D., Worrall, S., & Nebot, E. (2020). Efficient statistical validation with edge cases to evaluate Highly Automated Vehicles. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems*(doi: 10.1109/ITSC45102.2020.9294590), 1-8.

Leveson, N. G. (2011). *Engineering a Safer World.* London, UK: MIT Press.

New York Post. (2022, Sep 16). *Food delivery robot casually rolls through Los Angeles crime scene* . (CNN) Retrieved Feb 03, 2024

Ohno, T. (1988). *Toyota Production System* . New York: Productivity Press.

Serrat, O. (2017). Harnessing Creativity and Innovation in the Workplace. In O. Serrat, *Knowledge Solutions.* Singapore: Springer.

Smith, W. (2008). Making the Invisible Visible. *International Journal of the Spanda Foundation, 6*(1).

Smith, W. (Director). (2022). *LESI SIG Workshop Part 1 - Advances in Systems Sciences* [Motion Picture].

Smith, W. (2023). *What are the AIC Maps?* Retrieved 10 09, 2023, from https://powermap.odii.com/intro_signup.php

Smith, W. E. (2016). Purpose and Power in the Evolution of Socio-technical Systems Design . *Co-Creating Humane and Innovative Organizations*(Global STS-D Network Press September 2016).

Wei, J., & al, e. (2022). CoT prompting elicits reasoning in large language models. *Advances in Neural Information.* New Orleans, Louisiana, USA, SBN: 978-1-7138-7108-8.