The Safety-Critical Systems Club Newsletter

Safety Systems

Vol 32 No. 1 – Feb 2024



A MATTER OF TRUST
Embracing a future with AI

GLOBAL POSITIONING
How the world is planning to govern AI

DOUGHNUT RECIPES
Economic models for a safer world

Safety Critical Systems Club Supporting Development of Safer Systems
The Systems Safety Community and Club
SCSC

SCSC

thescsc.org


AI
Safety
Intelligence
development
evidence
management
Risk
Assurance
Systems
System
regulations
vehicle
Group
2023
2024
One
scsc
Safe
Approach
Artificial
need
also
human
use
Data
case
Time
project
however
economic
ensure
design
standards
engineering
environment
Software
reliability
Economics
Risks
issues
regulatory
Autonomous
Membership
Seminar
technology
Working
Doughnut
Automotive
analysis
control
potential
model
guidance
Critical
Newsletter
Introduction
Components
services
responsible
Helen
engineer
product
address
Don't
May
help
people
accessed
developed
models
Safetyrelated
Developing
Regulation
review
concept
planning
course
Performance
framework
understanding
verification
operational
highly
ethical
learning
specific
processes
vehicles
challenges
behaviour
Service
Part
made
article
across
therefore
Online
Website
Writing
although
problems
required
important
following
changes
failures
conditions
proposed
arguments
two
growth
harm
component
World
using
current
principles
role
Lead
work
used
even
due
take
many
key
training
level
must
2
engineers
Compelling
dependability
functional
held
John
including
making
Paul
good
Assurance
like
just
Risk
impact
4
Lead
Change
within
new
might
test
vehicles
way
environment
engineering
Software
reliability
makes
Regulation
2019
Developing
years
content
get
much
provide
information
cases
given
member
decisionmaking
another
technologies
confidence
projects
regulatory
Autonomous
Components
story
develop
responsible
Helen
services
engineer
product
address
Don't
May
help
people
Intelligence
accessed
developed
models
Safetyrelated
Developing
defining
2019
Regulation
review
concept
planning
course
Performance
framework
understanding
verification
operational
highly
ethical
learning
specific
processes
vehicles
challenges
behaviour
Service
Part
made
article
across
therefore
Online
Website
Writing
although
problems
required
important
following
changes
failures
conditions
proposed
arguments
two
growth
harm
component
World
using
current
principles
role
Lead
work
used
even
due
take
many
key
training
level
must
2
engineers
Compelling
dependability
functional
held
John
including
making
Paul
good
Assurance
like
just
Risk
impact
4
Lead
Change
within
new
might
test
vehicles
way
environment
engineering
Software
reliability
makes
Regulation
2019
Developing
years
content
get
much
provide
information
cases
given
member
decisionmaking
another
technologies
confidence
projects
regulatory
Autonomous
Components
services
responsible
Helen

Contents

WELCOME

Editorial 3

Opening words from the SCSC Newsletter Editor.

In Brief 4

Recent system safety news items from around the world.

FEATURES

Automotive AI: Help or Hindrance? 5

Helen Monkhouse asks how AI will impact the automotive industry.

If We Can't Define "safe", What Chance Does AI Have? 15

John Ridgway discusses what AI being "sufficiently safe" means.

AI Governance: Exploring Key Regulatory Approaches 21

George Mason and Greg Chance look at the state of global AI regulations.

Engineering, Ethics and 'Doughnut' Economics 29

Nick Hales suggests it is time to rethink our economic models.

REPORTS

A Brief Quest for Lesser-Spotted News Articles in Dependability 39

Wendy Owen draws insights from a range of recent news articles.

Seminar: How to write Compelling Safety Arguments 45

A seminar report covering the art of safety argumentation.

Seminar: What Makes Good Evidence for a Safety Case? 49

A seminar report on what makes good evidence for a safety case.

60 Secs with ... ChatGPT 61

ChatGPT answers some quick-fire questions on system safety and life!

GROUPS

Working Groups 65

Details of the SCSC Working Groups.

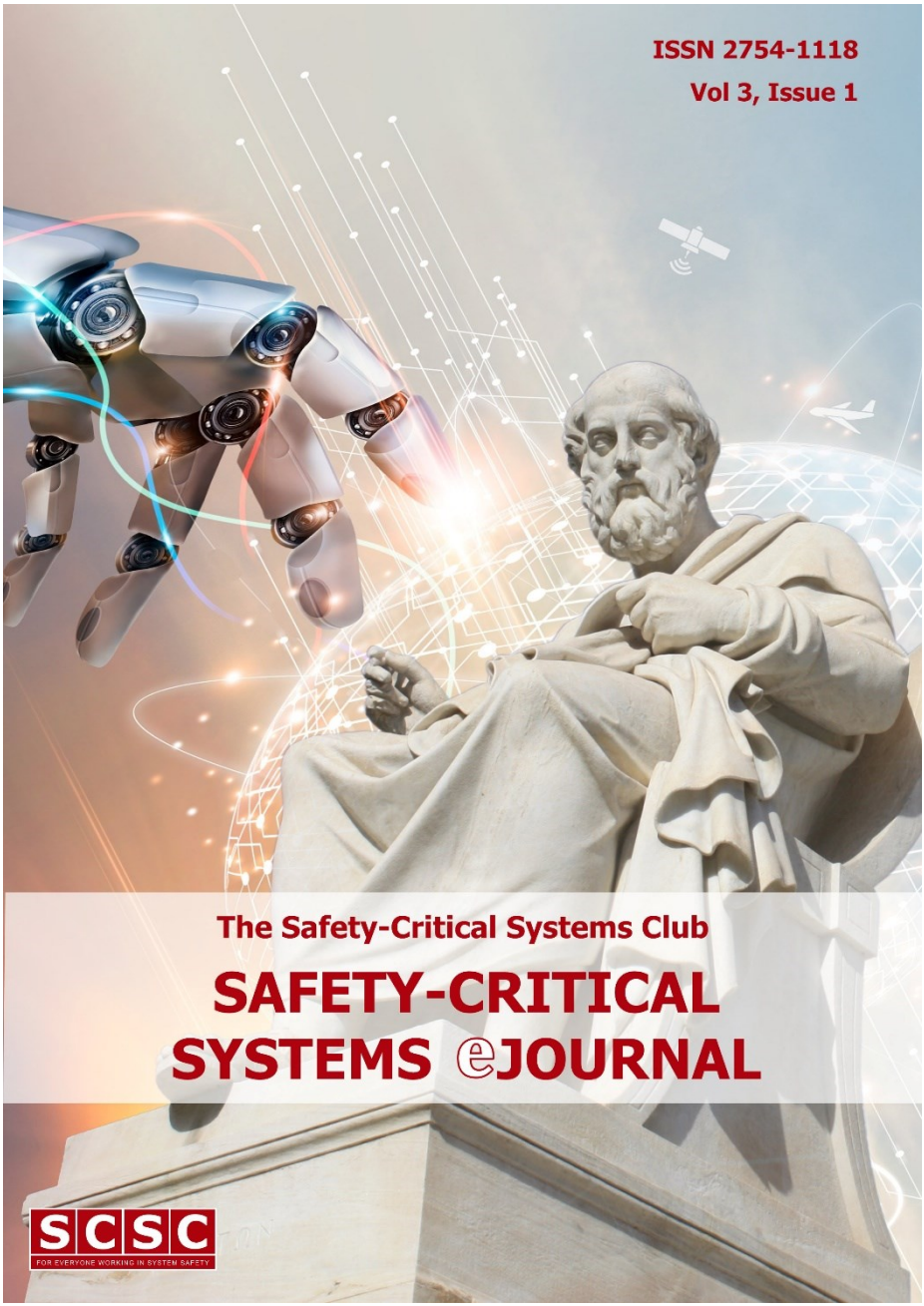
SCSC Steering Group 71

Who's who in the Steering Group.

EVENTS

Calendar 73

Events Diary 74


ISSN 2754-1118
Vol 3, Issue 1
The Safety-Critical Systems Club
SAFETY-CRITICAL SYSTEMS eJOURNAL
SCSC
FOR EVERYONE WORKING IN SYSTEM SAFETY

scsc.uk/r191

Editorial

If you find yourself routinely talking to a little box to set reminders, get the weather forecast, provide the best driving routes or generally enquire about anything on the internet, then you, like me and many others, are already embracing Artificial Intelligence (AI) and its influence seems set only to increase and indeed, is rapidly gathering pace.

Over Christmas, we were introduced to AI image generation applications, and my wife Heather and I have since been prolific users: she, to post amusing content on Facebook and for me, for example, to help generate the front cover of this newsletter. I was also interested to hear from my sister, a university lecturer, that AI is already embedded in her students' work with assignments stating what level of AI generated content is permitted. If our family experiences are anything to go by then AI is already taking a firm grip.



Applications that use Large Language Models (LLM) seem already extremely powerful and to illustrate, my 60-second interviewee is with ChatGPT and the results are, I hope you'll see, pretty remarkable, not least its predilection for extreme sports (as imagined by Heather's contribution shown opposite!), in fact, what perhaps gives ChatGPT away is how perfectly formed the answers are – lacking, in a sense, human imperfection.

With such clear utility, there does seem a clear and present danger that the adoption of AI will outpace our ability to govern its safety assurance, not least because we don't know how to do that yet as our traditional assurance methods are difficult to apply to AI. It is comforting however to note that the impact of AI is generally being treated seriously on several levels. In their article on AI governance, George Mason and Greg Chance discuss how various governments around the world are establishing laws for AI's adoption and at a grass-roots level, it was interesting to see the Royal Institution Christmas Lectures focusing entirely on AI and aimed at the 11-17 age group; effectively the generation that will be the main inhabitants of the AI world we create and custodians of its inventions.

The British Computer Society's Brian Runciman made an important statement in the society's Dec 2023 edition of "IT Now" magazine:

"If we want our generation's contribution to AI's evolution to be written about well by tomorrow's historians, we need to work just as hard at diversity as we do at AI engineering itself".

I wholeheartedly agree with this; there are already many examples of AI's learning developing bias such as excluding women applicants for historically male-dominated jobs and safety issues are lurking with, for example, known issues with gender and ethnic biases in medical devices.

I am therefore pleased to announce the formation of a new SCSC Working Group "Safe AI" that will hopefully provide a systems approach to the challenges of AI. Further joining details will be provided on the SCSC website in due course.

Paul Hampton
SCSC Newsletter Editor
paul.hampton@scsc.uk



In Brief








Post Office and Fujitsu chief executives face grilling by MPs on Horizon scandal



The chief executives of Post Office and Fujitsu are set to be questioned by MPs over the Horizon scandal.

The Horizon scandal saw more than 700 subpostmasters and subpostmistresses handed criminal convictions after faulty Fujitsu accounting software made it appear as though money was missing at their branches. Victims have since described being shunned by their communities, financially ruined and having their families destroyed.
independent.co.uk

Alaska Airlines grounds 737 Max 9 planes after section blows out mid-air



Alaska Airlines says it has returned some of its grounded Boeing 737 Max 9 jets to service after one of its planes lost a section of fuselage mid-flight.

The airline apologised to those affected, saying an inspection found "no concerning findings". The plane had to make an emergency landing in the US state of Oregon after an outer section of the plane fell off.
bbc.co.uk

IT failures causing patient deaths, says NHS safety body



"Urgent action is needed to address NHS computer failings which are causing harm to patients", the patient safety watchdog has said. The watchdog has evidence of patient deaths due to IT system errors. The government called the reports "concerning" and said it would work with NHS England to take necessary action to protect patients.
bbc.co.uk

Japan Airlines: Coastguard plane not cleared for take-off, transcripts show



A coastguard aircraft involved in a fatal collision at Tokyo's Haneda airport was not cleared for take-off, Japanese authorities have said. The smaller aeroplane collided with a Japan Airlines (JAL) passenger jet on the runway on Tuesday.

Released transcripts of air traffic control instructions just before the crash confirmed the JAL Airbus A350 was cleared to land.

Permission for take-off was not given to the coastguard Bombardier Dash-8.
bbc.co.uk

Automotive AI: Help or Hindrance?



In this article Dr Helen Monkhouse discusses the impact that introducing Artificial Intelligence (AI) into automotive products might have. Seen as an enabling technology for highly automated driving and driver support systems, AI also poses a disruptive challenge to vehicle safety assurance, so will it turn out to be a help or hindrance to the continued safe development of vehicle systems?

Travelling in an early 20th Century motor vehicle was a dangerous pursuit, and with vehicles being heavy and difficult to control, serious accidents were common. Since those early days road safety and vehicle safety have both seen a gradual evolution. During the early 1900's windscreen wipers, rear view mirrors and laminated windscreens were fitted to vehicles. The 1950's saw the introduction of 'passive' safety systems, including seat belts and the first airbags, that sought to minimise accident severity. 'Active' safety systems, like Anti-lock Braking Systems (ABS) and Electronic Stability Control (ESC), that support the driver to prevent a hazardous situation becoming an accident, are now standard.

The complexity of vehicle safety features continues to grow, with Advanced Driver Assistance Systems (ADAS) providing 'tactical' safety to the driver. That is, preventing driving situations becoming critical through the introduction of features, such as Automatic Emergency Braking (AEB), blind-spot monitoring and Lane Keeping Assistance (LKA). By their nature, 'tactical' safety systems require a comprehension of the vehicle's operating environment, or proper situational awareness, to operate correctly. It is for ADAS, Automated Driving (AD), and ultimately driver-less vehicles, where AI is seen as the enabler that will help propel these vehicle features into series production.

Although AI, and more specifically machine learning (ML), enables sophisticated vehicle perception systems to be realised, as a technology it possesses properties that differ from those technologies upon which existing safety practices are based. This means that AI technology is potentially a hindrance to the continued safe development of vehicle systems.

To unpick this help or hindrance question, I reflect on AI's use in automotive products, before then considering how this impacts product safety assurance. To keep article size manageable, I am ignoring the use of AI in automotive engineering development tools. Having considered how AI might impact automotive product safety assurance, I then reflect on what aspects of the classical design paradigm still have utility when AI technologies are deployed into safety relevant product. I conclude by returning to the question – Automotive AI: help or hindrance?

What is AI and where is it being used?

Artificial Intelligence

The International Standard on Artificial Intelligence (AI) ISO/IEC 22989:2022 defines an AI system as an "engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives." The standard also notes that the "engineered system can use various techniques and approaches…to develop a model to represent data, knowledge, processes, etc. which can be used to conduct tasks," and that AI can "operate with varying levels of automation" [1]. The creation of such a rational agent is concerned with developing machines capable of achieving goals and deliver the best outcome in the presence of uncertainty.

The theoretical techniques that underpin AI are not new. If we think about control theory, business operations or statistics, then it is typical to minimise or maximise to some function. It is high performance computing that now makes the use of such theoretical techniques tenable for industries like automotive. However, a potential limitation of an objectives led approach is the need to fully define the objectives [2]. For a self-driving vehicle, superficially the objective might be to move the passengers safely from Point A to Point B. However, when considered against external factors like, errors by other road users, bad weather or component failures, achieving strict adherence to this safety goal may well result in the AI based route planning function choosing never to route the vehicle out of its garage!

"Achieving strict adherence to this safety goal may well result in the AI based route planning function choosing never to route the vehicle out of its garage!"

AI use in automotive applications

As mentioned in the introduction, it is ADAS and AD perception systems where AI is being deployed within safety relevant automotive product. Side-stepping the issue of safe AI, by developing traditional procedural-based vehicle perception systems has limited utility. By simply considering the data representation of pedestrian forms in an Operational Design Domain (ODD), one can see that to fully define a pedestrian, such that it can be recognised by a procedural algorithm, becomes an intractable task.

The introduction of AD vehicle features that allow the driver to relinquish control of the

driving task to the automation, for a time, introduces a requirement for driver monitoring. Typically, camera-based systems within the passenger compartment monitor the driver's eye gaze and eye opening to assess alertness and fatigue. Decisions regarding whether to continue in AD mode, or when to hand-back control to the driver may be predicated on the assessment of the driver's current attentiveness. For example, conditional driving automation systems, like BMW's Personal Pilot or Mercedes Benz's Drive Pilot, allow the driver to potentially engage in non-driving task for several seconds while the feature is operational.

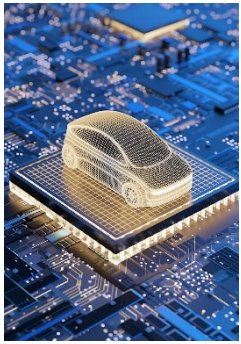
Perhaps not obvious to the end user, but AI based techniques are being employed in control adaptation too. Battery management systems that use adaptation techniques to improve the performance and longevity of electric vehicle batteries are in development, and adaptation techniques that improve system robustness, are widely used in engine and transmission control. Although typically deployed in the main control path, and not as part of the safety mechanism, the feasibility of safety monitoring adaptation is also being explored [3].

Automotive safety assurance

Although a complex system of systems, many of the engineering principles that underpin the development of the modern passenger vehicle are based on classical mechanical engineering principles. In this classical design paradigm, the vehicle comprises several discrete independent systems (what ISO 26262 refers to as an Item [4]). These Items deliver vehicle functions like power steering control, braking and stability control, that exhibit predictable behaviour that remains consistent throughout the vehicle's life. Today's new-vehicle homologation process is predicated on this property. That is, having verified that a new vehicle type meets the regulatory performance targets, assuming manufacturing quality is controlled and maintained, then the performance of the millionth vehicle produced will still adhere to those same regulatory targets.

Items like those described above will typically comprise an electronic control unit, having several sensor inputs and one or more output actuators. Requirements can be written that fully describe 'what' a given Item needs to do, from which the design specification can be created detailing 'how' the system implements its requirements. Assuming the Item's intended behaviour (the 'what') is not in itself unsafe, then functional safety is achieved if the architectural design prevents random hardware failures causing the Item's actuators to lead to harm, and if sufficient rigour has been applied during development to avoid systematic design errors.

Achieving safety of the intended function (SOTIF) becomes more involved for systems that sense their environment. The International Standard ISO 21448 describes the process of achieving SOTIF, in which system limitations (called functional insufficiencies) and the conditions under which these limitations may manifest (called triggering conditions) are identified, understood and managed [5]. A difficult task made more difficult when the system contains an AI component, because the way in which the AI model learns from the data is at a different level to the inferences made by the human analyst. That is, humans will typically analyse the system semantically, considering scene attributes like lighting conditions and the classification and position of objects, while the AI model operates on the underlying data.



The classic design paradigm assumes that components experience binary failures. That is, the fault is either present and remains present or it is not. For hardware components, the component tolerance could drift, or an open / short-circuit fault could occur. This will typically be because of aging and follow an exponential distribution. Software component failures are not caused by aging. Instead, a software systematic fault (i.e., software 'bug') can lay undiscovered, but will manifest itself every time the input conditions are such to generate the failure. If we think of an AI component as being an executable element (see Figure 1) then an AI component can also experience failures due to hardware random aging and software systematic failures, because of the hardware and software platforms on which the AI model runs. However, in addition to the classic random and systematic failures, AI components can exhibit misclassification errors.



For example, a ML vision system failing to classify an object as a pedestrian. However, given the relative complexity of the input space, the shape of the failure probability distribution is unknown. Potential methods to address these uniquely AI related faults are discussed below.

'Tactical' safety systems, whose correct and safe behaviour is predicated on correctly perceiving the vehicle's surroundings, challenge the classical design paradigm. Not only does fully defining the 'what' under all conceivable operating conditions become problematic, but the use of AI techniques like Deep Neural Networks (DNNs), makes understanding 'how' the system achieves the 'what' less explainable too.

Is the automotive product safety process broken?

It would be easy to dismiss the use of AI components in safety relevant applications, but this could be to the detriment of the potential benefits that deploying the technology might bring. To entertain using AI in a safety related application requires understanding the constituent parts of the AI component, to assess the impact that each might have on system safety, and in doing so devise means by which to mitigate potential errors.

Before delving into the impact that an AI component might have on safety, it is important to be explicit about what an AI component comprises. For the purposes of this discussion let us assume that to deploy an AI model, like a DNN, as executable code into an automotive application we need to create an Embedded AI System.

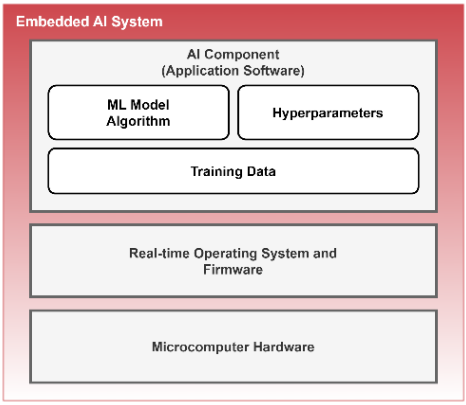


Figure 1 Embedded AI System

The Embedded AI System architecture is depicted in Figure 1. For the purposes of this discussion the architecture includes application software, an operating system, hardware specific firmware and computer hardware. Of relevance here is the application software, or AI component, which comprises

the model, hyperparameters[1], and training data. Supporting processes, not included on the diagram, but pertinent to the discussion include data management, the training approach (e.g., supervised learning, reinforcement learning) and test strategies.

How can an AI model cause harm?

At a superficial level the propagation of an AI model fault, through to unsafe vehicle behaviour can be represented by the diagram in Figure 2. Depending on its origins, the unsafe vehicle behaviour experienced can be classified as malfunctioning behaviour [4], hazardous behaviour [5], or loss of controllability due to misuse.[2] Although these classifications are unhelpful to the vehicle users experiencing the unsafe vehicle behaviour, they do illustrate how the different automotive standards dovetail together to address potential safety relevant AI faults.



Figure 2 AI component error propagation

From a safety perspective, techniques to prevent systematic faults in source code (e.g., divide by zero, incorrect variable names) and random hardware faults (e.g., component open or short-circuit) are addressed by the automotive functional safety standard ISO 26262:2018 [4]. ISO PAS 8800, a new international standard in development, seeks to address sources of functional insufficiencies in the AI technology capable of affecting vehicle safety [6].

"An AI component can be only as good as the model and the data that creates it"

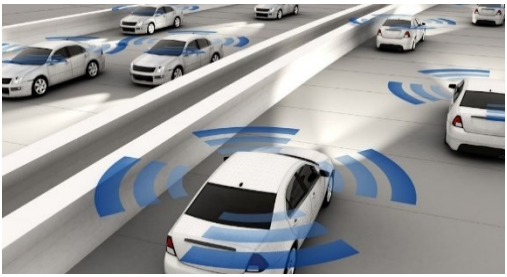
An AI component can be only as good as the model and the data that creates it. Errors in the specification, collection, generation and classification of data used to train and test the AI model can all result in functional insufficiencies, which can then contribute to a vehicle system error. Training data having missing data sets can lead to the misclassification of objects in the field-of-view. That is, semantic properties of the intended ODD may not be represented in the training data. In addition, the distribution of training data might not give a good approximation of the real-world characteristics, or it may shift over time. For example, if new car type or colour is introduced after the AI model has been trained and introduced into the field. The brittleness of DNNs also has the potential to impact safe behaviour. Common input

[1] Hyperparameters are the external parameters used to specify the AI model architecture. For example, the number of nodes and layers in a neural network, or the number of branches in a decision tree.

[2] Specifically, ISO 21448 refers to this type of misuse as Reasonably Foreseeable Indirect Misuse (RFIM) [4]. Examples of which could include, the driver falling asleep or unbuckling their seat belt and climbing into a passenger seat while an AD system is responsible for vehicle control.

perturbations like signal noise, weather conditions and objects appearing in atypical orientations all have the potential to produce output misclassification [7]. Even with correctly distributed data, a well-documented training error for DNNs is 'overfitting'. This can cause a DNN to give a corner case a wrong high-confidence classification.

In traditional system safety development lifecycles, Validation and Verification (V&V) both play an important role in building confidence in the system prior to deployment. When referring to V&V we perhaps immediately think of physical testing, however a significant part of the verification activity is system verification by analysis. That is, verifying that both the system architecture delivers 'what' is specified by the safety requirements, but also 'how' the safety architecture implements the safety requirements does not introduce new mechanisms for harm. A verification challenge for systems deploying AI models is the potential complexity of the model's architecture. This presents a scalability problem, as it is not uncommon for the number of parameters within a DNN to be into the millions.



It is a widely held view that for a system to be safe it must also be secure. Before passenger vehicles became networked, system security was a matter of physically preventing access to sensitive components, for example, using anti-tamper screws. High profile security breaches [8] have demonstrated the potential vulnerabilities that are introduced once vehicles are networked. However, by following cybersecurity engineering best practice [9] these types of vulnerabilities can be managed. The situation becomes more complex with the introduction of AI. Adversaries no longer need to gain access to the system in order to exploit that system's vulnerabilities. Instead, an adversary attack can exploit system vulnerabilities simply by manipulating the vehicle's environment. For example, by applying stickers to objects in the environment, or by partially obscuring a vehicle sensor.

The techniques that remain applicable

Irrespective of the technology, the underlining philosophy of safety engineering is to understand the causes of harm and then to devise mechanisms to prevent, avoid or mitigate that harm. What does change are the conditions under which risk should be subjectively assessed and therefore the integrity to which the product should be engineered. The existing automotive risk model [4] assumes that the driver is integral to vehicle control and ultimately responsible for vehicle safety. Risk mitigation under this paradigm involves deploying safety mechanisms that either reduce accident severity or increase the driver's controllability to a point where any residual risk would be rated as ASIL QM[3] [4]. For AD, part of the SOTIF process involves defining an acceptable residual risk [5]. For example, the number of persistent false negative pedestrian misdetections permissible by a DNN in a 30 frames per second video sequence. Defining acceptance criteria for an AD vehicle feature is both non-

[3] The ISO 26262 risk model classifies risk by considering accident severity, the probability of the hazardous event, and the driver's ability to control the hazardous event. The lowest level of risk is defined as QM – a level of risk where applying quality management principles are sufficient to achieve safety.

trivial and not uniquely an AI problem. As AD feature familiarity, regulation, and societal acceptance matures over time then best practice criteria will likely develop for the different features introduced.

The ability to shut-down a faulty system, reliability concerns and cost sensitivity of automotive production has meant that redundant safety architectures, such as a 2 out of 3 voting systems, are uncommon. However, AD availability requirements are increasing the need for redundancy. As are the introduction ML perception systems, where inputs from various sensor technologies can be fused together to improve overall system robustness.

Understanding how faults propagate and contribute to vehicle level hazardous behaviour, and ultimately an accident, remains important for systems that include AI components, as does the context in which the hazardous behaviour occurs. For example, the safety consequences of an AEB perception system's failure to detect a pedestrian far from the vehicle is less than the failure to detect a pedestrian located directly in front of the vehicle. In addition, plausibility checks can help the system to gain confidence in a prediction. For example, it is unlikely that a genuine object will disappear from the field-of-view within a few milliseconds. Instead, one would expect to see a real object (e.g., a pedestrian) remain in a video image for multiple frames.

I have long felt the Functional Safety Concept (FSC) to be the most important ISO 26262 work product, because, if written well, it conveys a logical description of how the item achieves safety. As a minimum the FSC should contain a description of the preliminary architecture (logical), the safety mechanisms, and the warning and degradation strategy. For a system containing an AI component, conveying the concepts used to achieve safety is paramount, as safety concepts will likely embody complex acceptance criteria and plausibility considerations. For high integrity systems ISO 26262 recommends, modelling the FSC in a language such as SysML [4]. This remains true for systems containing AI components. However, given the potential complexities and subtleties involved I feel that when the system contains an AI component, then prose should accompany the SysML model, to provide a written narrative of the concepts deployed.

Verification and Validation

In the 'How can an AI model cause harm?' section above, I discuss the importance of verification by analysis, both for achieving functional safety and SOTIF, but suggest that the introduction of AI components makes verification by analysis difficult if not untenable. That said, system level verification by analysis and indeed semantic analysis has an important role to play in AI safety, and that is regarding robustness. Using guideword led analysis (e.g., HAZOP), semantic factors having the potential to affect the sensor system (e.g., windscreen smudges, dirt, blur) can be considered systematically. Once identified, these artefacts can be included into test data, thus improving data completeness. Confidence in the AI model's robustness can be further improved by using classic design of experiment techniques to test input sensitivity. That is, exploring the AI model's sensitivity to the input space by changing input magnitude ranges or by using different combinations of inputs [6]. However, devising a sensitivity measure is a non-trivial task.

Validation remains a challenge for systems containing AI components, particularly testing the unknown or the unpredictable, given that is not possible to extrapolate from known to

unknown test points. For systems containing an AI component, validation refers to the activity of checking that the AI system implements the system safety requirements, therefore it is typically done at vehicle level by the vehicle integrator (e.g., vehicle manufacturer). The focus for the safety validation is achieving full test coverage through the systematic exploration of all relevant test scenarios, and automotive standards now exist dedicated to scenario-based testing [10].

Monitoring vehicle in-field performance as part of a Cybersecurity Management System (CSMS) is now a requirement [9, 11]. Broadening the scope of in-field monitoring to include the identification of new or useful data, can help maintain an AI component's performance through runtime learning. By detecting and sending back out-of-distribution data to a central data centre, the level of uncertainty (i.e., the degree to which the data is out-of-distribution) can be assessed, and ultimately new data sets can be derived and new AI models created. These new AI models can then be put through the normal development cycle and distributed to the vehicles through over the air updates. Allowing in-the-field self-learning may be the utopia, however, from a validation perspective self-learning is considered too problematic and therefore infeasible in the short to medium term.

"Allowing in-the-field self-learning may be the utopia, however, from a validation perspective self-learning is considered too problematic and therefore infeasible in the short to medium term."

In conclusion – help or hindrance?

As I have discussed, a prerequisite for today's 'tactical' vehicle safety features is their ability to obtain and maintain situation awareness. However, given the complexity of the ODD, the ability to define the design space such that traditional procedural algorithms can be used for vehicle perception systems becomes an intractable problem. By virtue of their capability to solve problems (e.g., pedestrian detection) which cannot be fully specified (e.g., characterising all pedestrians within the ODD), deploying AI based systems helps solve the system perception problem.

As we have seen, the introduction of AI systems into automotive product clearly challenges the status quo, by making many of the assumptions that underpin the achievement of functional safety and SOTIF obsolete. Therefore, we could jump to the conclusion that introducing AI components into automotive product is a hindrance to the continued achievement of vehicle safety. However, as this discussion and standards like ISO PAS 8800 highlight, many of the system safety practitioner's existing tools and techniques remain relevant. With less safety requirements typically being defined for the AI component than for a traditional item, the focus of product development effort will need to change. For example, focusing effort into devising test strategies, capable of demonstrating AI model robustness and safety.

As a closing thought, I believe there is a credible route to series production for AI based systems. For me that involves articulating the safety concepts deployed clearly and coherently within the FSC and conveying the safety confidence (with uncertainty) achieved through a comprehensive safety case argument.

References

[1] International Organization for Standardization (ISO), "BS EN ISO/IEC 22989 Information technology – Artificial intelligence – Artificial intelligence concepts and terminology," 2023.

[2] S. Russell, P Norvig, "Artificial Intelligence, A Modern Approach" Pearson, 2022, Fourth Global Edition, ISBN 1-292-40113-3.

[3] F. Botes, et al., "Development and Optimisation of an Adaptive Safety Monitor," SAE WCX 2018.

[4] International Organization for Standardization (ISO), "ISO 26262 Road Vehicles – functional safety" ISO, 2018.

[5] International Organization for Standardization (ISO), "ISO 21448 Road Vehicles – Safety of the Intended Functionality" ISO, 2022.

[6] International Organization for Standardization (ISO), "ISO PAS 8800:2023 Road Vehicles – Safety and artificial intelligence" ISO, 2023, CD draft.

[7] O. Willers, et al., "Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks," Cornell University, 2020.

[8] C Miller, C Valasek, "Remote Exploitation of an Unaltered Passenger Vehicle", academia.edu, 2015.

[9] International Organization for Standardization (ISO), "ISO 21434 Road Vehicles – Cybersecurity engineering" ISO, 2021.

[10] International Organization for Standardization (ISO), "ISO 34502 Road Vehicles – Test scenarios for automated driving systems – Scenario based safety evaluation framework" ISO, 2022.

[11] UNECE, "Regulation No. 155 - Uniform provisions concerning the approval of vehicles with regards to cyber security and cyber security management system," 2021.

Image attributions:
top image: © Shutterstock id: 589267616
object detection: © HORIBA MIRA
car on chip: © Shutterstock id: 2187915949
networked vehicles: © Shutterstock id: 700959994

Dr Helen Monkhouse, HORIBA MIRA Ltd.

Dr Helen Monkhouse is a Chartered Engineer and system safety practitioner who has worked in the automotive domain for over 25 years. She is currently a Functional Safety Chief Engineer at HORIBA MIRA. Helen is a UK Expert to the British Standards AUE/32/1 Safety and Security of EE Components in Road Vehicles committee, and a co-author of MISRA's Guidelines for Automotive Safety Arguments.

The author retains copyright of this article.

TechTrip: Visit and Guided Tour of the Shuttleworth Collection

THE SAFETY-CRITICAL SYSTEMS CLUB, TechTrip:

Visit and Guided Tour of the Shuttleworth Collection

Saturday 20 April, 2024 - Biggleswade, Bedfordshire

The next SCSC TechTrip will be to the Shuttleworth Collection (https://www.shuttleworth.org/) on Saturday 20th April 2024.

This is a fabulous collection of historic aircraft and vintage vehicles in lovely grounds. The aircraft are all still airworthy!

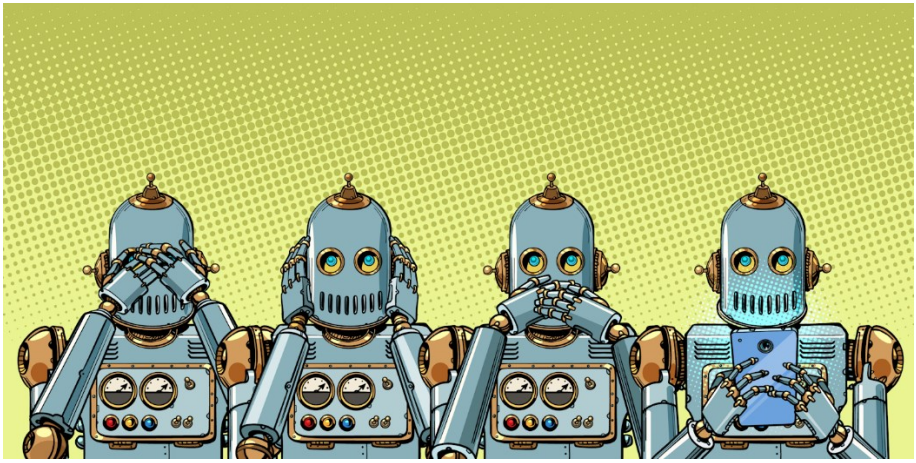
https://youtu.be/_FD7SJ8MNoE

We meet at 1pm and have a guided tour of the hangars followed by free time in the gardens and house.

Cost is £25 pp.



If We Can't Define "safe", What Chance Does AI Have?



John Ridgway discusses the challenges of designing autonomous Artificial Intelligence decision-making systems to be "sufficiently safe". As "sufficiency" requires a value judgement in the safety-engineering domain, he suggests that it is more "artificial wisdom" that is necessary, which will require AI to operate within a suitably encoded ethical framework. But how would such a framework be encoded, on whose value system would it be based and how much should emotion play in decision making? John explores these questions and provides some insights into tackling these issues.

It's All About Values

I don't think I have to remind any of the readers of this newsletter that safety is a value-laden concept. The roots of this problem are essentially epistemic. Such is the nature of uncertainty – that we cannot say with one hundred percent confidence that any system or situation will prove to be safe. Indeed, safety engineers learn at a very early stage in their professional development that the word 'safe' should always be accompanied by the modifier 'sufficiently'. We accept that sufficiency is all we can achieve, and within that acceptance can be found the value judgements. These judgements take many forms and are an expression of intelligence; or, to be more precise, wisdom. In fact, one can define wisdom as the judicious application of limited knowledge, as constrained by an advantageous ethical framing. So when thinking about the safety of AI, my first question would have to be this: How can we ensure that any machinery designed to make autonomous decisions does so within a suitably encoded ethical framework? Or, to widen that question, even when AI is only playing a supporting role in human decision-making, how can we stop AI from facilitating unethical intentions?

Those who foresee a problematic future often assume that the critical point will only be reached once AI has achieved a level of intelligence that surpasses our own. Such concerns may make for a good science fiction movie, but I believe superior intelligence is not really the main issue, and for two reasons.



Firstly, there is what I would call the baby-sitter syndrome. When we place another intelligent agent in charge of something we hold precious, we do so on the assumption that the surrogate shares our values. And just to make sure, we instruct them regarding those values, using natural language. The issue is whether or not such instructions are an effective encoding of the required values and whether they have been understood. We don't normally become overly troubled if the baby-sitter has superior intelligence to our own, as long as that intelligence is used to understand the imperative for respecting the terms and conditions of their employment. The dynamics of the problem are no different with artificial 'baby-sitters', but the practicalities are fundamentally different, not least because the use of natural language as machine instruction to encode ethical values (such as respect) remains a technical challenge.

Secondly, as with all system safety issues, the problems do not lie solely within the inherent properties of the system, but rather within the relationship existing between the system and both its operational environment and its human operators. In particular, safety risk often lies in the gap lurking between how reliable the system is and how reliable it is assumed to be. That gap can exist at any level of system intelligence. One can argue that such problems already exist at the human/AI interface and can be discerned in the degree to which we are often unable to even reliably identify the workings of AI. And even when we are fully aware, we have a tendency to anthropomorphize and thereby bestow upon AI capabilities and sensibilities that it doesn't actually possess. Furthermore, AI drives expert systems and we all feel a lot happier when we see the trappings of expertise – it's as if AI wears a white coat and has a stethoscope draped across its shoulders. So it is not the sophistication of AI that is the problem so much as human gullibility and complacency. Artificial Intelligence had the capacity to make fools of us long before it could possibly gain the motivation and skill to do so. It may also be worth keeping in mind that AI is already a tool of human rivalry.

The Machine Intelligence Research Institute

It is at this stage that I think it is only fair that I admit, when it comes to talking about AI, I too have the capacity to deceive. I am no expert but I flatter myself that I have the guile and sophistication to pass myself off as one. I became aware of this 'talent' some years ago when I was approached by Luke Muehlhauser of the Machine Intelligence Research Institute (MIRI) to take part in an interview [1] that would be posted on their website. He had become aware of my work writing for the SCSC Newsletter and had, from that, deduced that I had expert credibility. Indeed, he initially referred to me as Dr Ridgway (I assure you that I have no such credentials; I dropped out of a PhD in theoretical nuclear physics the day after it occurred to me that it was quite a difficult subject). Flattered and alarmed in equal measure, I obliged to take part but was very careful to make as my opening statement, *"I am not an expert and I would not wish to pass myself off as one".* The rest of the interview passed amiably enough but I have to say that if Luke had hoped to gain anything that would help

his institute better understand the challenges of AI safety, he must have been sadly disappointed. Indeed, I suspect that the real legacy was my own heightened interest in the works of MIRI.

If you visit the MIRI website [2] you will see that it is a charity announcing itself with the following statement:

"MIRI's mission is to ensure that the creation of smarter-than-human intelligence has a positive impact. We aim to make advanced intelligent systems behave as we intend even in the absence of immediate human supervision."

They continue:

"MIRI focuses on AI approaches that can be made transparent (e.g., precisely specified decision algorithms, not genetic algorithms), so that humans can understand why AI systems behave as they do. For safety purposes, a mathematical equation defining general intelligence is more desirable than an impressive but poorly-understood code kludge".

Anyone with a safety engineering background would readily recognise the need for transparency. Nevertheless, the idea of "*a mathematical equation defining general intelligence*" is intriguing to say the very least. Those wishing to learn more about MIRI are invited to visit its website, but for the purposes of this article I would like to draw your attention to one MIRI paper in particular, written by professors Nick Bostrom and Eliezer Yudkowsky, and titled '*The Ethics of Artificial Intelligence*' [3].

Establishing An Ethical Framework

The MIRI paper concentrates upon the characteristics one would require of any AI algorithms that are used to take over social functions, and draws the conclusion that responsibility, transparency, auditability, incorruptibility and predictability are primary amongst them. Astutely, the authors acknowledge that "*this list of criteria is by no means exhaustive*". Which is just as well because, although it is a similar list to that required for any safety-related system, it does not appear to address the need for a clearly specified, algorithm-based policy to inform risk-based decision-making. For example, from where would the AI algorithm derive its bases for the minimisation and prioritisation of risk, cost benefit calculations, or the transfer of risk between stakeholders? These are all value-based considerations that would require more than the number-crunching wizardry of an idiot savant.

We don't just need "*a mathematical equation defining general intelligence*"; we need a mathematical equation defining general wisdom.

Delving further into the problems of predictability, the paper refers to the concept of Artificial General Intelligence (AGI) and sees within that concept a clear technical difficulty with the safety of such systems. As the paper puts it:"

"It is relatively easy to envisage the sort of safety issues that may result from AI operating only within a specific domain. It is a qualitatively different class of problem to handle an AGI operating across many novel contexts that cannot be predicted in advance."

Indeed it is a qualitatively different class of problem, but the paper is still not focusing upon the more intractable problems that arise when one considers the ethical dimensions of decision-making – the demands of generality will just make those problems worse. This is because, to achieve generality, an AI algorithm will need to be adaptable both in comprehension and purpose without transgressing an overarching ethicality. In other words, it is required to exhibit what I would call 'Artificial Wisdom'. As a consequence, we don't just need *"a mathematical equation defining general intelligence";* we need a mathematical equation defining general wisdom.



But where is this wisdom to come from? Are safety engineers to play God and fashion AI systems in their own image? Do we ourselves have a sufficiently good track record in meeting expectations of sufficient safety? Can we even agree upon the definitions of relevant terms, such as 'reasonably practicable'? How good are we in dealing with novel situations in our profession? Have we got to grips with all of the moral dilemmas that are encompassed in safety engineering policy? And if we are not on top of all of this, how are we to expect autonomous AI to operate safely and wisely, given that its only steer will be coming from the way we have encoded within it our own moral basis for decision-making – assuming, of course, that such a machine encoding were to be possible?

The Nature Of Human Decision-Making

As a final concern, I would like to turn to another aspect of human decision-making that could present a challenge for those who would wish to replicate it in autonomous safety-related systems. It relates to the role of emotion and is encapsulated within a question asked by the neuroscientist, Robert Sapolsky: "*Do we mostly think or feel our way to deciding what is right?"* [4]



For a long time, philosophers and scholars of human cognition thought that emotion was a potential impediment to good decision-making. Plato, for example, used the metaphor of the charioteer struggling to keep two winged horses under control. One of the horses represented the emotional and impulsive side to human nature, whereas the charioteer represented the rational soul, seeking to override and control such impulses. Modern neuroscience, however, has led to an understanding that emotion is not so much a potential nuisance as it is an essential component in making decisions.

Neuroscientist Antonio Damasio reports [5] that a patient with damage to an area to the brain vital in the mediation of emotional states (i.e. the ventromedial frontal cortex) was not left, as one might expect, with the capacity to make purely rational decisions. Instead, he was left with an obsession with irrelevant detail and an inability to act decisively where it matters. His intelligence was not impaired but he acted fecklessly, as though lacking intelligence. Sapolsky reports that other patients with similar brain damage express opinions in laboratory experiments that are objectively rational but nonetheless demonstrate an alien moral compass; for example, it is better to sacrifice one's own family than a greater number of total strangers. Contact with one's emotionality, it seems, is one of wisdom's vital ingredients.

One can debate why this should be so, but one has to presume that there is evolutionary advantage to such an arrangement. Whether or not this provides a good argument for basing artificial intelligence upon a similar model, i.e. one with an interplay between artificial rationality and artificial emotion, is a question that I'm sure can be better answered by those more qualified than I. But I would like to raise the following possibility: If ethicality and morality are at least partly founded upon emotionality, then the possibility arises that a machine's instantiation of ethical decision-making may also have to employ some form of analogy to the role of emotion in human decision-making. And if this were to be necessary, how could it possibly be achieved? What would that do to the mathematical equation defining general wisdom?

If ethicality and morality are at least partly founded upon emotionality, then the possibility arises that a machine's instantiation of ethical decision-making may also have to employ some form of analogy to the role of emotion in human decision-making.

Here's Hoping

I have written this article for no better reason than to share with you the musings of a non-expert. Artificial Intelligence is a highly technical subject deserving a better qualified appraisal. However, it is also a concept that invites a great deal of philosophical debate and conjecture. We are warned that the stakes are very high and that miscalculations at this stage may have very serious consequences. This seems plausible, if only because many safety applications have been mooted and, even without a dominance of AI, safety considerations present the engineer with many challenges that go well beyond the simple objective of ensuring system reliability. We can look forward to embracing AI and all of its advantages but we must be wary of the prospect of it embracing us. If we expect it to keep us sufficiently safe, then we probably need to do a lot more to determine what that even means.

References

[1] https://intelligence.org/2014/03/08/john-ridgway-on-safety-critical-systems/, accessed Nov 2023

[2] https://intelligence.org/research/, accessed Nov 2023

[3] https://intelligence.org/files/EthicsofAI.pdf, accessed Nov 2023

[4] "Behave: The Biology of Humans at Our Best and Worst", Chapter 13: "Morality and Doing the Right Thing, Once You've Figured Out What That Is", ISBN 9780-0-09957-506-1.

[5] "Descartes' Error", Antonio Damasio, ISBN 978-0-099-50164-0.

image attribution:
do no evil: Illustration 143452766 | Communication © Valeriy Kachaev | Dreamstime.com
robot holding baby: AI generated Midjourney
face: Illustration 289198092 © Golib Tolibov | Dreamstime.com
winged horses and chariot: AI generated Midjourney

John Ridgway, retired

John Ridgway is a retired traffic systems consultant who led the software development team responsible for the M25 MIDAS computer control system. Subsequently, he undertook a functional safety management role whilst working for a contractor commissioned to provide a computer control system for the Highways Agency's M42 Active Traffic Management system.

The author retains copyright of this article.



Seminar: Frameworks for Safe AI Systems

THE SAFETY-CRITICAL SYSTEMS CLUB, Seminar:

Frameworks for Safe AI Systems

Thursday 25 April, 2024 - London, TBA

This seminar is the first of three in 2024 on an AI theme. This one will look at the ethical, legal and regulatory frameworks and societal/social aspects of AI systems. It will also consider high-level goals and safety objectives for AI.

Speakers include:

Dai Davis - Percy Crow Davis & Co Ltd

Zoe Porter - AAIP, University of York

(The second and third seminars will look at construction, implementation and operation of AI systems.)

Your Ad Here!

Do you have a product, service or event you would like to advertise in the Newsletter? The SCSC Newsletter can reach out to over 4,000 individuals involved in Systems Safety and so is the perfect medium for engaging with the community.

For prices and further details, please get in touch with the Newsletter Editor.

AI Governance: Exploring Key Regulatory Approaches



George Mason and Greg Chance look at the state of global AI regulations. They review the characteristics of key regulatory approaches that have been proposed and discuss their strengths and weaknesses as well as the future of AI regulation.

Artificial Intelligence (AI) has experienced considerable growth over the past decade, primarily due to the significant advancements made possible through deep learning techniques. The current generation of AI markedly outperforms its predecessors, being capable of a diverse range of tasks that previously were exclusive to skilled humans. At some activities, AI can even exceed human performance.

As AI continues to become proficient at human tasks and those beyond human capabilities, there is mounting concern about how it may impact society. Specifically, some worry that AI may be detrimental to humanity. This has led to an increase in demand for *trustworthy* AI that retains the benefits of AI whilst minimising its risks.

The concept of 'trustworthy' AI is loosely defined, with numerous definitions of varying degrees of formality appearing across the literature. Furthermore, alternative words such as 'responsible', 'assured', 'ethical' and 'safe' are sometimes used in place of trustworthy. Regardless of the exact term used, common characteristics that appear in definitions for trustworthy AI include fairness, transparency, robustness, accuracy, accountability and security.

To ensure the use and development of trustworthy AI, jurisdictions worldwide have begun drafting AI regulations that outline rules and principles designed to instil trust in AI systems. In this article, we delve into the regulatory landscape for AI, beginning with the motivation for such regulations, followed by an analysis of the characteristics of key regulatory approaches and then a discussion of their respective strengths and weaknesses and the future of AI regulations.

The Need for Regulations

Until recently, the risks associated with AI were minimal or only applicable in specific circumstances: AI had limited capabilities, meaning there was less criticality, and even the most capable forms of AI typically required specialist equipment and expertise to use.

Presently, however, AI can perform numerous tasks at a human level and is even accessible on consumer-grade equipment by people with little or no expertise. These factors significantly increase the potential risks that AI poses to society, including:

- Mass unemployment – As AI continues to master new abilities, human jobs are becoming increasingly vulnerable. Unlike other technological advances throughout history, the flexible nature of AI to learn an ever-widening range of skills could make it challenging for people to find new roles in the face of redundancy.
- Privacy violations – AI can be used in ways that violate privacy; for example, mass surveillance through facial recognition systems. Moreover, the use of large amounts of data to train AI systems creates the potential for mishandling and leakage of personal information.
- Bias and discrimination – Improperly trained AI systems can exhibit biases and discriminatory behaviours. For instance, when AI is used in the recruitment process, unjust hiring decisions may be made based on factors that should not affect the selection procedure.
- Physical or psychological harm – AI systems installed within physical environments, such as self-driving cars, could behave dangerously in situations where the AI encounters unseen data. Additionally, excessive reliance on tools like chatbots could lead to issues such as social isolation and a weakening of critical-thinking skills.
- Social unrest – Generative AI techniques can synthesise disruptive yet false content. For example, deepfakes that depict notable figures presenting misinformation or chatbots that can produce convincing yet fabricated information. This problem is magnified as the content propagates across social media.

Concerns such as these and more motivate the need to regulate the use and development of AI systems.

Key Regulatory Approaches

The terrain of AI regulations is rapidly evolving. The exact timeline for when the first regulations were proposed is unclear, but most key efforts have occurred within the past three years. As interest in trustworthy AI continues to grow, it is expected that new regulations will be created whilst existing proposals undergo regular revisions.

Notable efforts towards AI regulation have been made by the UK, EU, China, Brazil and Canada. Other jurisdictions such as India are also engaged in discussions on the development of AI regulations; however, their progress is less advanced. Interestingly, the US has

not yet proposed a comprehensive set of federal regulations for AI; although, it has introduced numerous laws and guidelines relating to this field.

Except for China, to date, no jurisdictions have yet finalised and enforced regulations at a national or international level. The following is an overview of those regulations that are most mature.

UK



In March 2023 (last updated in August 2023), the UK published a whitepaper outlining its proposed AI regulations [1]. Being so recent, they may be subject to significant revisions. A key priority of the regulations is to foster innovation amongst AI developers by minimising the burden of stifling legislation.

The regulatory framework is principle-based, encompassing five key principles:

1) safety, security and robustness
2) appropriate transparency and explainability
3) fairness
4) accountability and governance and
5) contestability and redress

To ensure the longevity of the regulations' relevance, the whitepaper provides the broad definition of AI as being a system that is adaptive and autonomous.

The regulations are sector-based, involving existing regulatory bodies to enforce the principles appropriately for each sector. Instead of detailing penalties for noncompliance within the regulations themselves, it falls to the relevant regulatory body to determine appropriate sanctions as needed.

The regulators tasked with enforcing the principles will vary by sector; at present, it remains uncertain exactly who will be responsible for each area. However, it is likely that the Digital Regulation Cooperation Forum (DRCF) may take an overarching role in this regard.

EU



Of all AI regulations still being drafted, the EU's are closest to being finalised. First proposed in April 2021, the drafting process is nearing its conclusion; the European Parliament intends to begin their enforcement in 2024 [2].

A key aim of the EU regulations is to uphold human rights. To achieve this objective, it adopts a tiered, risk-based approach that categorises AI systems based on their potential disruption to society into four levels of risk: unacceptable, high, low and minimal. Furthermore, the regulations empower those negatively affected by AI to seek redress.

The regulations' definition of AI is based on the one given by the Organisation for Economic Cooperation and Development (OECD). In contrast to the UK's definition, this definition is more detailed, specifying behaviours and techniques that characterise an AI system.

Enforcement of the regulations will occur at both national and international levels. Nationally, each member state of the EU is responsible for enforcing the regulations within their country; the specific regulatory bodies who will be responsible for enforcement may vary by country.

Internationally, enforcement will be overseen by the European Artificial Intelligence Board, comprising a representative from each member state. The regulations outline specific financial penalties for noncompliance, with a maximum penalty of up to 35 million euros or 7% of annual worldwide turnover.

China



As of the time of writing, China is the only jurisdiction to have implemented AI regulations at a national level: in March 2022, it enforced its regulations on the management of algorithmic recommendations [3]. Since then, China has continued to develop and enforce further regulatory measures.

The motivations behind Chinese regulations differ somewhat to other jurisdictions. China has an ambitious goal to become the world leader for AI development and these regulations represent a substantial step in this direction. Furthermore, they are important for managing the creation and dissemination of content with respect to ensuring national interests.

A further difference in the Chinese regulations is that they do not appear to provide a general-purpose definition of AI. Instead, they list specific technological behaviours relevant to the purpose of each regulation; this may encompass technologies not traditionally considered to be AI whilst potentially excluding others broadly recognised as such.

Despite these differences, the regulations do have some similarities to regulations from other jurisdictions, such as stipulating that AI systems are accurate, fair and safe.

The regulations are enforced by the Cyberspace Administration of China, with penalties for noncompliance seemingly financial in nature.

Brazil



Brazil has proposed various AI regulations dating back to 2019; however, these were met with limited approval from Congress. The most recent proposal, which replaces previous ones, was introduced in May 2023 [4]. It shares many similarities to the EU's, from which it likely drew inspiration.

Similarities include using the OECD's definition of AI as a basis and adopting a risk-based approach. Additionally, both proposals prioritise human rights; however, the Brazilian regulations seem to emphasise the importance of this even more so by making it central to their regulations.

The main enforcement body for the regulations is yet to be determined; however, it is likely to be the Brazilian Data Protection Authority (ANPD). Penalties for noncompliance vary, including warnings; fines of up to 50 million Brazilian reals or 2% of annual revenue; publicising the infringement; and prohibition or limitations placed on the offender's ability to do further AI development.

Canada



Like Brazil, Canada's approach to AI regulation (proposed in June 2022) shares certain similarities to the EU's; in particular, prioritising human rights and defining AI based on the OECD's definition [5].

In common with the UK's proposed regulations, Canada follows a principle-based approach, emphasising the importance of transparency, accountability, data protection and ethics whilst minimising disruption to innovation. It is possible that the UK was inspired by Canada when drafting their regulations.

Enforcement of the regulations is likely to be centralised, with the task falling to The Minister of Innovation, Science and Industry in conjunction with an AI and Data Commissioner. Violations of the regulations will incur financial penalties of up to 25 million Canadian dollars or 5% of total revenue.

US



The US has not yet proposed a comprehensive set of federal regulations for AI; although, some regulations have been implemented at a state level. So far, the US has primarily approached AI governance through the development of individual laws tailored to specific contexts.

However, in October 2023, the White House published an executive order on the safe, secure and trustworthy development and use of AI, outlining eight guiding principles for the creation of trustworthy AI systems [6]. Although the executive order is not a comprehensive set of regulations, it suggests that one may eventually follow.

Discussion

AI regulations are still a new concept, with only China having finalised and enforced any such regulations at a national level. All other regulatory efforts are still in the drafting stage.

Among those under development, the EU's regulations are closest to completion; they are expected to be enforced during 2024. These regulations will likely prove highly influential on the development of other AI regulations worldwide. Indeed, it seems that they already have been instrumental in the development of Brazil's regulations and may have served as inspiration for Canada's.

The UK aims to strike a balance between ensuring trust and promoting innovation amongst developers.

The UK's regulatory framework presents an interesting variation compared to others. Whereas most other regulations primarily focus on ensuring that AI is trustworthy, the UK aims to strike a balance between ensuring trust and promoting innovation amongst developers. China's regulations also stand out due to their unique blend of trustworthiness-focussed aspects and provisions aimed at ensuring national security by managing the creation and distribution of content.

Although the US has not yet drafted any regulations, given their global prominence in the development of AI and the issuance of an executive order focussed on the need for trustworthy AI, it can be expected that federal AI regulations are likely to be created in the future.

A consideration that does not seem to be addressed by the regulations is how they are to be enforced at an individual level, such as someone using their personal computer within their home. With the proliferation of open-source models and the ready availability of tools and equipment to work with AI, combined with the abundance of free educational resources

available to learn how to develop and use AI models, it is relatively straightforward for anyone with a modest budget and an internet connection to engage in AI use and development. Whilst the personal use of AI poses fewer risks compared to an organisation or government using it (for instance, mass unemployment is an unlikely risk), numerous other risks persist, such as the capacity to generate misinformation or harm mental wellbeing. It remains unclear how the proposed regulations aim to address this issue; it appears that the current regulatory proposals focus primarily on the use and development of AI by organisations and governments.

Another factor to consider is the effectiveness of the specified penalties for noncompliance. Where they are defined, sanctions are typically in the form of fines; however, if an organisation can enjoy substantial profits from noncompliant AI systems and these profits outweigh any financial penalties, even significant fines may not serve as a deterrent due to the net profit after fines still being greater than that generated by compliant AI systems. In this regard, the UK's regulations may be at an advantage, since penalties will be determined based on what is deemed appropriate within a given sector. This approach could potentially inspire other jurisdictions to follow suit if blanket financial sanctions are found to be ineffective.

The definition of AI provided within each set of regulations is another area that may undergo revision. Definitions range from broad in scope to specific and detailed. Each approach has its advantages and disadvantages: broader definitions can potentially apply to future forms of AI which may function significantly differently to current ones; however, it could become unclear to stakeholders and regulators what types of technology qualify as AI without clear delineation. Stricter definitions have the opposite advantages and disadvantages. It will be interesting to observe how effective each definition proves in practice.

An intriguing question is which regulations are likely to become dominant on the global stage; which, if any, will become the global standard? Factors that could influence this include trade, leadership in AI development, political influence and the proven track record of effectiveness of the regulations. In the current geopolitical climate, however, it is unlikely that a single approach will emerge as dominant; instead, it seems likely that multiple regulations will become established and coexist for the foreseeable future. Stakeholders will need to be cognizant of these different regulations if they wish to conduct business in those jurisdictions.

An intriguing question is which regulations are likely to become dominant on the global stage; which, if any, will become the global standard?

A final point to consider is that regulations represent only one aspect of the strategy to ensure trustworthy AI. Other aspects include the development of tools and techniques to assess the risks of AI systems and to ensure they satisfy the criteria for trustworthiness. Although considerable progress has been made in this area through continuing research efforts, much work remains to be done. Similarly, further work is needed to crystallise the process of enforcing the regulations; specifically, identifying the regulatory bodies responsible for this task and outlining the procedures that stakeholders must adhere to in order to fully comply with the regulations.

Conclusion

Regulation of AI is still in a nascent stage and its future remains uncertain. However, it seems likely that the EU's regulations will take a pivotal role moving forward. Despite the uncertainties surrounding their effectiveness, these regulatory efforts demonstrate a serious commitment from both experts and legislators worldwide to ensure that our future – where AI is expected to hold significant influence – is one of prosperity and not disaster.

References

[1] Department for Science, Innovation & Technology (2023). A pro-innovation approach to AI regulation. Great Britain. (CP 815). ISBN: 978-1-5286-4009-1.

[2] European Commission (2023). The Artificial Intelligence Act. [online]. https://artificialintelligenceact.eu/ Accessed 11th January 2024.

[3] China Law Translate (2022). Provisions on the Management of Algorithmic Recommendations in Internet Information Services. [online]. https://www.chinalawtranslate.com/en/algorithms/ Accessed 11th January 2024.

[4] OECD (2023). Brazil's path to responsible AI. [online]. https://oecd.ai/en/wonk/brazils-path-to-responsible-ai Accessed 13th January 2024.

[5] Innovation, Science and Economic Development Canada (2023). Artificial Intelligence and Data Act. [online]. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act Accessed 13th January 2024.

[6] The White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. [online]. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ Accessed 14th January 2024.

Image Attribution
Clarote & AI4Media / Better Images of AI / User/Chimera / CC-BY 4.0
flags: 20631874 © Jameschipper & 162596679 © Fourleaflover | Dreamstime.com

George Mason • Frazer-Nash Consultancy Ltd

George Mason is an engineer at Frazer-Nash Consultancy, working in the Trustworthy AI team within the Digital Systems Assurance group, where his work focusses on the risk assessment and safety assurance of AI in safety-critical systems. Prior to this, he was a member of the Trustworthy Adaptive and Autonomous Systems and Processes research team at the University of York, where he used formal methods to develop an approach for assuring safety in reinforcement learning. He was awarded a PhD for his research, which won the Best PhD Project award at DCAART'17. His continued research interests lie primarily in the development of trustworthy AI techniques and systems.

Greg Chance • Frazer-Nash Consultancy Ltd

Greg Chance is a consultant in the Digital Assurance Group at Frazer-Nash Consultancy and an Honorary Fellow of the University of Bristol. His interests are AI assurance and methods to develop responsible AI and verification and assurance techniques. He has previously held academic positions researching robotics and AI and has worked on numerous InnovateUK and EPSRC projects about assuring safety of AI systems, including autonomous vehicles, UAVs, soft robots and swarms. He has also developed testing methods for simulation-based verification of AI, including agent-based reinforcement learning and crowdsourcing.

The authors retain copyright of this article. All opinions expressed in this article are entirely the authors' own and do not reflect the views of their current or previous employers.

Seminar: Developing Safe AI Systems

THE SAFETY-CRITICAL SYSTEMS CLUB, Seminar:

Developing Safe AI Systems

Thursday 27 June, 2024 - London, TBA

This seminar is all about how to develop and build safe AI systems: it will discuss the methods, tools and techniques that should be employed, the all-important approach to testing and V&V, and also the quality and safety assurance processes that govern what has to be done. The speakers will outline possible ideas and approaches for these evolving areas.

This seminar will be held in central London, venue TBA.

Speakers include:

Richard Hawkins, University of York - *"Through-life assurance of ML systems (TBC)"*

Further details TBA.

Engineering, Ethics and ‘Doughnut’ Economics



Nick Hales introduces “Doughnut Economics”, a relatively new economic model that is intended to address the gulf between the preoccupations of mainstream economic theory, which relies on growth economics to address developmental issues, and the growing real-world crises such as global inequality and climate change with disrupters like Artificial Intelligence (AI) making the landscape ever more complex. Nick goes on to suggest a modification to recognise that greater cooperation is required amongst the underdeveloped, developing and developed nations.

The Mission Statement of the Systems Approach to Safety of the Environment Working Group (SASEWG) is, “To produce clear guidance on how engineered systems should be developed and managed throughout their entire lifecycle (including through system or environment changes) so as to preserve, protect and enhance the environment.” [1].

The Rules of Conduct of the IET [2], include the following two rules (14 and 15 respectively):

- Members shall at all times take all reasonable care to limit any danger of death, injury or ill-health to any person that may result from their work and the products of their work
- Members shall take all reasonable steps to avoid waste of natural resources, damage to the environment, and damage or destruction of man-made products

These ethical viewpoints have to be balanced by affordability, but there is a disconnect. It is much like the disconnect children have between the new born lamb they wish to hold close at a 'petting' farm and the lamb they have for Sunday lunch. In a world of limited resources, it is suggested that we cannot continue to look away from the problems of impact of projects simply because it is not what we like to do. In engineering, there is always the pressure to accept "the bottom line", money. The cost/benefit analysis, (CBA), is of course an essential for all projects, but there are hidden costs that are eating away at the health of the planet so, both now and in the future, we must examine the larger picture and work to protect the well-being of humanity and biodiversity, the fates of which are intertwined.

This article looks at the problem, in the modern world, of continuing to use economic models developed in the past when the issue of development was assumed to be solved by growth economics alone. A relatively new economic model, Professor Kate Raworth's "Doughnut Economics" [3], is introduced. The intent of the model is to encourage the closing of the "visible gulf between the preoccupations of mainstream economic theory and the growing real-world crises such as global inequality and climate change", and a modification to that model, developed by the author, is illustrated and the logic and reasoning behind it given.

Somehow, we must create:

- Acceptable resource usage cycles which work with rather than exploit nature and people
- Processes to utilise the resources that do no permanent or long-term damage
- Products developed by the processes applied to the resources, which must truly hail themselves as eco-friendly products in use and in disposability

A good reason to explore 'Doughnut' economics is that it can be growth agnostic. It avoids the dogma that pits ecological issues against growth obsessions.

Note: An analysis of how an engineering firm would use the model proposed by this article will be given in a later issue of this newsletter.

A Brief Overview of the Effect of Extant Economic Models

There is a tendency to conflate economics with greed. An example of this conflation, used ubiquitously, is Adam Smith's words "It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest." [4]. The problem is when self-interest becomes massively disproportionate to need. Aristotle distinguished between 'chrematistics', the art of acquiring wealth, and 'economics' , derived from the Greek for household management and could be used as the basis for caring for the natural world in the management of projects, looking after our 'home' planet.

To ameliorate the worst aspects of the potential future economic problems, some economists are now suggesting that there are alternative ways of looking at economies. Even the greatest economists have had their doubts about whether their subject is a science. One thing both Friedrich Hayek and John Maynard Keynes agreed on, best summed up by Hayek, was that "an authority, which in economics no man ought to possess" [5] is conferred by the Nobel Prize. We should therefore be open to new thinking, open-mindedness being a quality of good engineers.

Raworth's analyses show there are 2 conundrums:

Firstly:

- No country has ever ended human deprivation without a growing economy
- No country has ever ended ecological degradation with a growing economy

Secondly:

- We have an economy that needs to grow whether or not it makes us thrive
- We need an economy that makes us thrive, whether or not it grows

Solutions to these conundrums are urgently needed. Raworth proposes alternatives to seven given here in Table 1. There are of course many more.

Table 1: Economic Models and their Doughnut Economic Equivalents

Current Economic Model	Doughnut Economic Model
GDP success measure	DOUGHNUT safe space for humanity
Self-contained economy	Embedded economy
Rational economic man	Social adaptable humans
Mechanical equilibrium	Dynamic complexity
Growth will level up wealth	Distributive by design
Growth will clean it up	Regenerative by design
Growth addicted	Growth Agnostic

The Subtlety of Harm Due to the 'Bottom Line'

It is hard to avoid causing harm in one way or another when a project gets underway and this section illustrates typical problems, depending on the project, extracted from an article in Green World [6]. Table 2 gives types of problems and some examples.

Table 2: Human Activity and its Environmental Impact

Activity	Examples of Problems
Agriculture	As with palm oil plantations in Malaysia and Indonesia, destroying orangutan habitats.
Introduction of invasive species	Such as goats in the Galapagos, eating the plants the tortoises depend on.
Logging	Including the loss of tropical forests, (and many species with them), for furniture.
Construction	As with turtles on Turkish beaches with nowhere to nest now or distracted by lighting.
Climate Change	Warmer conditions can drive endemic rare species farther up their mountain homes for cooler air until food runs out at the tree line. These conditions are also bleaching corals where many fish breed.
Pollution	As when chemical fertilisers create huge harmful algal blooms in marine environments.
Human intrusion and disturbance	For example international travel has spread the fungi that are killing off amphibian species worldwide.
Fishing	Commercial boats trawl the ocean floors and catch whole fish shoals, including the young, leaving none to breed in adulthood. Climate change bleaching corals where fish breed. Now planning to destroy the ocean floor with trawls where life and death has been undisturbed for millions of years.
Water Management	Diverting rivers for agriculture can wipe out habitats and damming rivers can prevent species migrating up river to mix gene pools or breed.
Transportation	Building roads can enable poachers to find valuable species.
Capitalism	The system can perversely raise the price of rare species as each one shot reduces the numbers and increases the rarity, putting the price up further.

One example of how complex the prevention of extinction of a species can be concerns the sociable lapwing, which thrived on steppe grassland. That seems simple but the conversion of Ukraine and western Russia to wheat growing took most of these lands and the sociable lapwing prefers short grass for its nest and colonies of fellow lapwings. In addition, the rest of the steppe was home to grazing animals including the Saiga antelope but these were hunted to the edge of extinction so the grasses now grow long. The only remaining homes for this lapwing are close to villages where domestic herbivores keep the grass low. Unfortunately, that brings their nests and young into conflict with pet cats. Hence extinction becomes a possibility without due care.

The Doughnut Economic Model

Two schools of economics theory have dominated for decades. The first is that an economy is built the way Adam Smith imagined, by enterprise with selfish intent that spills over to others as it grows. The second is that sharing and a more co-operative existence will be the only way the earth and all its people can advance, as various manifestations of Marxist thought describe. Unfortunately, both thus far have led to the creation of new elites. The market economies are dominated by multi-millionaires and billionaires with their unwillingness to part with any of their earnings in the form of tax. The Marxist economies are dominated by brutal dictatorships, exclusive elitists and ideologically dogmatic committees.

In terms of technology, Artificial Intelligence, (AI), is certain to be influential where economics is concerned. Both these theories will seek to maximise the use of AI. Once the atomic bomb was created, efforts began to prevent the misuse of the devices. Equally, so must it be with AI. We are beginning to see a rapid growth in the use of the technology in many sectors and it is already recognised that this rapid expansion could lead to massive unemployment. Distribution of the wealth increasingly thus generated will become a big problem. Coping with a future without employment, poses a problem for governments and businesses. One solution a number of local governments are experimenting with is the concept of a Citizen's Income (CI) [7] which would be a sum of money to which all would be entitled, even if they were not part of the production cycle.

"AI is certain to be influential where economics is concerned. … once the atomic bomb was created, efforts began to prevent the misuse of the devices. Equally, so must it be with AI"

Into this battle of ideologies appears the relatively new Doughnut Economics model, (Figure 1), described for the lay person in [1]. The social foundation in the use of the doughnut model employed in this article, is the United Nations' 12 Sustainable Development Goals to achieve by 2030 [7]. It should be noted that this illustrated doughnut is not the only use of the model and an internet search will provide a library of custom variations on the theme.

It is important to understand that the doughnut economic model as presented here, is as a planning tool for engineering companies. It is designed to help answer the questions, "What is it we can do, in designing, constructing and maintaining this project that will help lift some above or toward the social foundation?" and "What can this project do to reduce the pressure on the planet due to ecological ceiling overshoot?" Some of the conclusions arrived at may be cost free or even reduce costs.

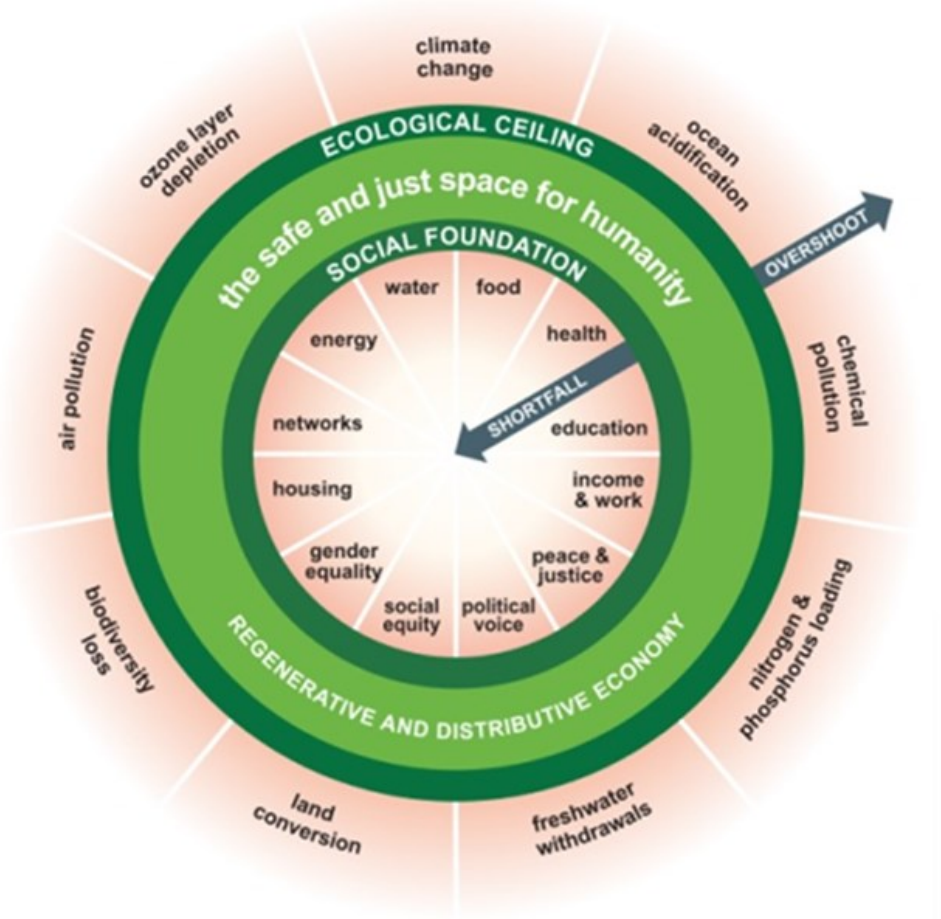


Figure 1: A Standard Doughnut Economic Model

In this doughnut, the 'safe' place for humanity into the future is represented by what would be the dough in a ring doughnut. As can be seen in the diagram, there are two of what one might call 'unsafe' places for humanity.

The first is outside the doughnut, crossing the ecological ceiling, where environmental damage occurs due to the various causes shown in the diagram. A projects intention should be to minimise those aspects of the project that are causal factors contributing to those identified effects, drawing the projects influence ever closer to the Ecological Ceiling until they pass through, where possible, that barrier into "the safe and just space for humanity". Being outside the doughnut is referred to as overshoot. To avoid overshoot, where possible, projects should introduce mitigations of their own and restorative activities to reverse the destructive aspects of previous projects. As an example, one could restore beavers to a habitat in order to create a flood plain, which beavers do naturally, to mitigate climate change effects on a new downstream eco-friendly factory construction project.

The second forbidden zone is inside the ring of the doughnut where the result of a project may lead to deprivation of one sort or another. One of the effects illustrated is a lack of water. If the water requirements of the project are such that it will deny sufficient water for the environment or people of any location it affects, then it has removed the possibility of a safe and just space for humanity. This is referred to as a shortfall. The project will have moved the local population across "the social foundation" barrier and away from the "safe and just space for humanity".

Should we be concerned if the world rejects the Doughnut model? The view of the author is that if someone is not concerned about the unnecessary deaths of millions of people every year, they are unlikely to be concerned about the extinction of any or most animal species. That would leave us, or more so our descendants, with the prospect of walking an earth with little of its beauty left and endless elitists arguments as to whether or not one should be concerned about inhuman politics?

Economic Solution Woes

There are consequences of mainstream political attempts to solve economic woes. A few examples are given here. All expressions quoted in the bullet points below are from Kate Raworth's book [1].

- Deregulate or re-regulate finance leading to "speculative bubbles, house price hikes and debt crises", bringing poverty and destructive desperation
- Promises to cut red tape that in doing so dismantle "legislation that was put in place to protect workers' rights, community resources and the living world"
- Turning public services such as hospitals, railways and water supply into "private revenue streams". These inevitably look to avoid any loss-making work in the former public stream such as long-term care on the social front, complex operations in medical care and unprofitable routes in the transport domain
- The approach espoused by Bjorn Lomborg, who famously has advocated in 'The Skeptical Environmentalist' [8], that the environment can be priced as "ecosystem services and natural capital, assigning it a value that looks dangerously like a price"

The result of the latter is that whoever wants to see tigers or a Javan rhinoceros in the wild, must come up with the money themselves, enabling those not bothered to dodge responsibility for wanton destruction of the planet known as 'Ecocide' [9]. A interactive map of the species that have gone extinct in the USA shows the astonishing lack of management of resources of the past two centuries [10]. This terrible attitude ignores what future generations may want. Extinction is forever and people alive today have no right to decide which flora and fauna are worthless. On such principles, even primary education could become available only to an elite. Slogans can always be devised by environmental crisis deniers. For Lomborg supporters, it has always been that people-focussed growth is more important than biodiversity or other ecological considerations. For Australians in the recent vote on aboriginal rights it was "if you don't know, vote no", even though the previous campaign claimed the rights were too prescriptive.

The "Three Rings" Solution

The author has developed the concept of three doughnuts as a development of that in Figure 1. The fundamental of the Three Rings solution and the necessity for such a methodology is that people in the developed nations would not support any suggestion of a leap to an egalitarian world, that led to massive debts, simply to satisfy critical voices considering only the underdeveloped countries situation. Exploitation has occurred in what is now the developed world. It was for the benefit of rich elites not the poor of those countries. Wealth for the people has emerged from those in power taking responsibility for democracy and human rights. What is needed is a good will between underdeveloped, developing and developed nations and it is proposed that the Three Ring Doughnut solution will enable this.

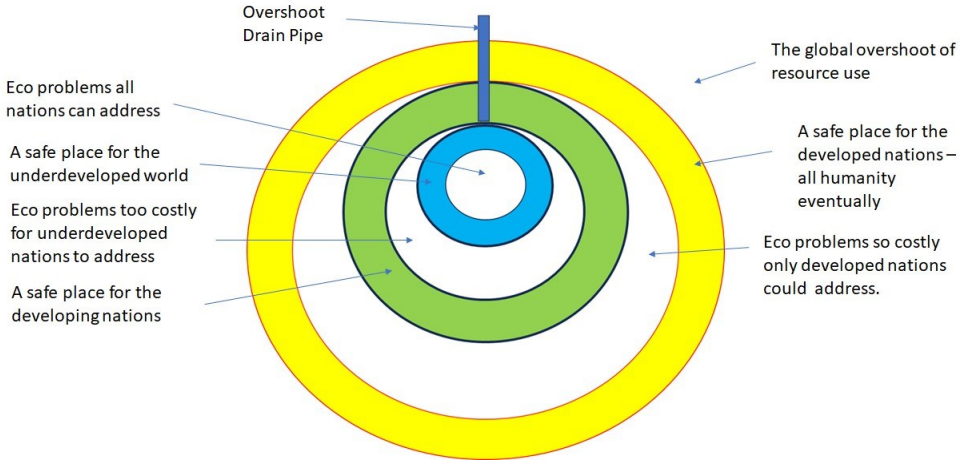


Figure 2. The Three Ring Doughnut

In the model, the developing and underdeveloped nation doughnuts only impinge at a tangent to the economically more powerful ring, or rings, within which they sit. This represents the responsibility to help nations with less powerful economies and political institutions develop; it is not reciprocal and poor economies should be able to depend on richer nations to help with environmental and social problems. The drain pipe in the diagram (Figure 2) illustrates that an underdeveloped country should consider some forms of pollution it creates as overshoot for all countries. The more highly developed the end-user, the more responsibility must be taken.

Ring Parameters for Different Applications

Each ring will have parameters in the planning of a project that differ from the parameters of other projects. Aspects of life of those affected by the project but not necessarily direct stakeholders can be improved by the project if enough thought is put into planning. It is important to orient benefits of the project for the best advantage of communities, especially if the project is in one of those countries identified as underdeveloped. The project should check to see what the people involved and the local economy are missing of the United Nations goals [7]. Next, put in measures and costs to ensure that at the very least nothing decreases the welfare of the people or is damaging to the environment. Such ethics may be highly beneficial to sales as people become more aware of the need for conservation.

A rough guide for a start might be to choose the inner third ring as representing those countries where the average wage is $100 or less per month. Developing countries could be exemplified with wages from $100 to $1000 per month, Malaysia being at the top end of that scale. Developed economies could then be defined as anywhere that monthly salaries are over $1,000 per month. None of these are definitive, each project will be different.

The Impact of AI on Economic Models

AI could be an unprecedented disrupter for human economics so it's important to consider how much impact AI could have on these models and how robust they are to the challenges and novelties posed by AI. I provide my views by considering some specific questions on AI:

Is AI just a tool to drive efficiencies helping the rich get richer or is it a significant boon to the environment by making human activities less wasteful of resources?

Our political choices will decide how much of the benefit is distributed. What Raworth noted about the market economy we can apply to AI, "Like fire, it is extremely efficient at what it does but dangerous if it gets out of control" [3]. Many disciplines have looked at complexity, including our own in safety critical engineering wherein we have to focus on anticipating emergent properties more than most. Raworth encourages embracing the complex in economics too, (see Table 1). Our grounding in complexity alerts us to incautious decisions that would be detrimental to the safe space for humanity. AI must obey and stay within the boundaries we set. The greatest benefit may then indeed come in the reduction of waste of money and time and thereby resources.

It is possible that AI will make seismic changes in employment profiles. How will that affect the economic models including the doughnut models?

There are now surgeon-controlled robots performing operations. The Royal United Hospital in Bath, declares their Da Vinci XI surgical robot will support more than four hundred complex surgical operations annually, saving time, resources and lives [12]. How AI affects current employment paradigms depends largely on whether or not the doughnut economy triumphs or the destructive aspects of the well-established historical economic models. Raworth [3] discusses two juxtapositions, self-enhancement counter-balanced by self-transcendence and openness to change counter-balanced by self-restriction.

A reasonably equitable future is possible with AI but where ever possible it must be applied to benefit people and planet, not to only provide wealth for those who already have more than enough. The key in Bath is that the hospital remains committed to providing for all citizens. Although currently under control by a surgeon, an AI computer control system would be able to make all the critical decisions leaving more of a hands-off, eyes-on approach to be taken by the surgeon, akin to airline pilots in modern digital cockpits. The RUH notes that the robot is removing health inequalities, a UN goal.

A future run by robots enslaving humans to protect the world may be science fiction, but can we envisage a world where there is an extreme reliance on AI in the future and where will that will place us economically?

Be cautious in development is the answer. The troubling questions of autonomy as yet unanswered are discussed in [13]. Where the economy is concerned there is undoubtedly the potential to impoverish billions with AI but also to enhance living for most people.

Conclusion – Don't be a Monkey

The Three Ring 'Doughnut Economics Model' is a non-egalitarian viewpoint at implementation. However, the three rings, with good will, should reduce to the two rings of developing and developed economies in a couple of decades. That is in stark contrast to the indifference that has been creating an ever-growing gap between rich and poor.

"Doughnut" economics is growth agnostic making it comprehensible to all. Using it as a tool will help identify procedures projects can engage in to help those affected by the project but not within the original specification. A researcher has discovered a group of macaques that have learnt how to use tools to open oyster shells [11]. The macaques choose the largest oysters because the rewards are greater. The oysters are now however endangered because of this selectivity. The future requires that to preserve our planet, we can continue to use tools but must resist being like the macaques. Are we intelligent enough to conserve and share our food and other resources?

More information about Professor Raworth's work and current action may be found at the Doughnut Economics Action Lab [14].

Image Attribution:
top image: AI art from Microsoft Image Creator
fig 1: https://commons.wikimedia.org/wiki/File:Doughnut_(economic_model).jpg
fig 2: Nicholas Hales (author)

References

[1] https://scsc.uk/ge accessed 29/11/2023

[2] https://www.theiet.org/about/governance/rules-of-conduct/ accessed 29/11/2023

[3] Kate Raworth "Doughnut Economics"

[4] Adam Smith An Inquiry into the nature and causes of the wealth of nations Volume 1

[5] Friedrick August von Hayek – Banquet Speech – https://www.nobelprize.org/prizes/economic-sciences/1974/hayek/speech/ - 1974, accessed 11/12/2023

[6] Nicholas Hales Green World Magazine Issue 87. www.greenworld.org.uk 2015

[7] UNDP Sustainable Development Goals https://www.undp.org/sustainable-development-goals#:~:text=The%20SDGs%20are%20a%20bold,conflict%20and%20climate%2Drelated%20disasters, accessed 11/12/2023

[8] A Citizens' Income: A Foundation for a Sustainable World – 21 Nov. 2003 - by Clive Lord

[9] Ecocide website https://www.stopecocide.earth/ accessed 03/01/2024

[10] https://extinctionmap.org/index.html, accessed 01/12/2023

[11] BBC documentary "Primates" episode 3

[12] https://ruhx.org.uk/having-an-impact/our-campaigns-and-projects/ accessed 04/01/24

[13] The SCSC Newsletter Vol 27 Number 1, "Autonomy's Troubling Questions" https://scsc.uk/r152:1

[14] DEAL https://doughnuteconomics.org/team accessed 03/01/2024

Nick Hales

Nicholas Hales, (C.Eng), is a retired critical avionics engineer and project manager in the MOD's Technology Office where his work involved de-risking technologies prior to full investment. He is currently a member of both the Data Safety and Systems Approach to Safety of the Environment SCSC working groups. He recently appeared as the Sultan of Morocco in an adult version of the pantomime Dick Whittington.

The author retains copyright of this article.

A Brief Quest for Lesser-Spotted News Articles in Dependability



In this article, Wendy Owen explores and draws insights from a range of recent newsworthy articles from the rail, aviation, communications and marine sectors.

If you keep up with engineering news, rarely a week goes by when one might spot or hear of an article in the news – online, on TV or radio, in a newspaper, through networks – that has safety and/or operability implications. We see them but, unless we have some direct or vague involvement, we often don't have time to dig further and just skim over them. I started paying more attention to news articles for a few months, purely to see what appeared and to ponder on them rather than skim over them.

One thing that has made me ponder over the last few years reading various articles is how on Earth certain untoward scenarios that occurred during operations were seemingly missed or omitted, presumably during the design stage. Such scenarios being not necessarily hazardous, but yet also seemingly highly unlikely to occur if good processes had been implemented during the design phase. Perhaps one might expect such anomalies to occur with very novel or very complex systems, but these are often not the subject of the reporting. In highly regulated sectors, safety cases are common, but dependability cases – in their own right – are not as common as one might expect. Even the word dependability tends to still be restricted to certain sectors. In this article, I am talking about dependability (with an overlay of resilience). You'd think dependability would be having a field day with current

advances in technology, with enthusiastic Reliability, Availability, Maintainability (RAM) engineers keenly demonstrating the great characteristics of various new-fangled innovative systems. But rarely does one see job adverts for RAM engineers nowadays, and reliability engineers apparently keep databanks up and running.

From what I've been seeing in the news, something is not right down in the reliability woods and beta factor coppices. I began to postulate what has been happening? In the case of highly regulated sectors, I wondered if in some cases HAZOPs have been morphing into HAZIDs and thus losing their OPerability quotient, or if the guidewords had not been sufficiently diverse to pick up future reliability issues, or if they had become more like tick-box exercises rather than exercises in broad and innovative thinking. Was any RAM analysis actually being conducted? I aimed to pay a bit more attention to news feeds and seek out articles where the OP, or perhaps a RAM or Human Factors (HF) analysis, may have helped prevent large outages, shocks, or accidents.

Over the course of a few months the following interesting examples appeared. I make no claim to knowing if any of the associated companies or projects here did a dependability, operability, HF or resilience analysis, these are purely my own observations as a third party.

Titan Submersible Implosion (June 2023)



Titan steering control – Logitech F710

I feel reasonably SQEP to talk about this one as I've worked on similar submersibles in the distant past, and a former manager was an advisor to the Kursk submarine accident. Consequently, when this happened, I was glued to the ongoing newsreels, mostly thinking if it hasn't imploded by now, it will do shortly, and commented as such to the UK head of naval safety, who was obviously keeping largely quiet on the matter until the situation was confirmed. In the end, the submersible clearly wasn't structurally reliable for its intended scope of operations, ergo neither was it safe. The investigation is ongoing at time of writing, but it has become evident that the company had not sought certification for the vessel, arguing that "*Bringing an outside entity up to speed on every innovation before it is put into real-world testing is anathema to rapid innovation*"[1], a subject in itself.



Titan end cap recovered

What I want to tease out here is that in this case it became obvious that 'successful' operations and the purported reliability of the missions to date (not many) had been prioritised over real design safety margins, and by a long chalk. The concepts of safety margins and structural reliability clearly went out the window - or weren't even in its frame. Issues were raised by the design team but ignored. And it hadn't been rigorously tested before going into operation.

The message here is just because it appears reliable, dependable, or even a fantastic innovation, doesn't mean it's safe, and it also doesn't mean you should ignore implementing a comprehensive reliability testing plan or creating a DRACAS.

Chaos on Sydney Trains (July 2023)

I have worked on a few Australian rail projects but not in Sydney, so of course this one caught my eye. Their metro is "fully automated" with Communications-Based Train Control (CBTC) signalling technology. A similar set-up can be found in Copenhagen, Milan and a few other places. This metro began operation in 2019 and carries about 20 million passengers annually. At time of writing, it has 13 stations and 22 trains with 22 miles of track, and over 30 stations under construction with aspirations of 70 miles of track. This metro proved very robust one day in July 2023 whilst the rest of the Sydney train network entered, well, a state of chaos. Even, a highly predictable state of chaos.



Ornate Sydney Station departures board (now in a museum)

"Vast swathes of Sydney's train network came to a standstill in the middle of the rush hour on Thursday because a single member of staff called in sick. Commuters were stranded on trains and on platforms across the city as the issue, which … had cascading effects on much of the rail network. … "Sydney Trains had late notice of resourcing issues this afternoon at the Rail Operations Centre. As a result, the continuity of the operations plan was impacted,"…."The resourcing issue involved highly specialised roles and the incident will be reviewed."….. By late afternoon, another signaller had come on shift. However, by that time, chaos has spread to trains throughout Sydney…. Commuters were given conflicting information and drivers had little idea how long the disruption would last. Enormous queues were seen at major hubs …. Replacement buses had to be scrambled to take passengers home…. every train line suffered some form of delay. Only the Sydney Metro, which is run completely independently from the Sydney Trains network, was unaffected. Passengers were left bemused as to why there was no backup in case just a single person couldn't come into work. One commuter said "It's absolutely crazy. How can one person calling in sick make such a massive difference?" [2].

And they were right. It is absolutely crazy that one person calling in sick could disrupt the journeys of tens of thousands or millions of people.

Maybe operability and human factors were considered in the course of design and planning, and this was just a case of 'bad luck' but, having spent a few years working for a signalling supplier/manufacturer, this particular event was akin to the loss of a critical path in project management or a single point failure in a physical system resulting in a high consequence event. It 'screams' of a dire lack of Human Factors Integration (HFI). Clearly, one member of staff being unavailable should not result in major signalling system downtime or major disruption on mainlines. This issue should have been picked up pre-commissioning.

The concept of "failure" applies as much to humans as to systems. If you're looking at

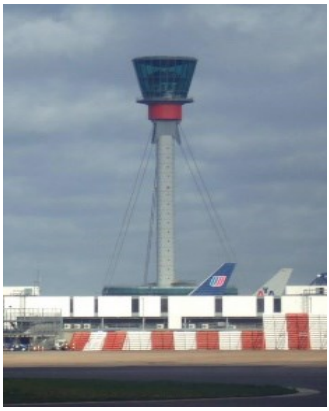
transport systems involving high interaction with the general public, or if throughput of people is critical, or if you need continuity of operators to operate (as in this case), then humans are integral and critical to successful (and reliable) operations and system availability.

The main messages from this one are please don't forget about the humans, including sick ones, and don't forget to do operability and HFI studies. My own experience says maintainability assessments can also reveal similar gaps, so don't forget those either.

NATS Goes Haywire (August 2023)

Having very briefly worked on NATS during a major upgrade about 20 years ago, I was surprised to read this article [3]. However, I won't delve deeper into this one as no doubt it has been so high profile that there will be lots of people here writing about it who know a lot more about the systems than me. Although it appears that the system had the correct and safe reaction in the circumstances i.e., it shut down, externally it appears as an operability issue – the transport system ground to a halt (pardon pun), similar to the Sydney trains.

It's a good example of where there is a question about how much unavailability or downtime are the travelling public prepared to accept, and how is that balanced against safety or company reputation?



Heathrow traffic control tower

Comms Chaos Down Under (Nov 2023)

Even worse than the train network chaos back in July, on 8th November a report came on the news about "Optus outage: Millions affected by Australian network issues" [4].

On closer inspection, *"The failure caused transport delays, cut hospital phone lines, and shut down payment systems… more than 10 million people and thousands of businesses were affected… Services were restored after around 12 hours of disruption….there was no evidence of a cyber-attack. The firm blamed a "technical network fault" and said more time would be needed to investigate the root cause…. People across Australia were unable to call emergency services and critical helpline numbers. Train services in the state of Victoria were also temporarily crippled."*



This serves as an excellent example of how widespread the impacts can be across industry and everyday lives of "uptime"/"downtime" and safety from a single source of failure in national critical infrastructure.

We can hope that the communications network concerned are having a serious investigation of the cause and the subsequent chain of events and are planning to employ a large department of systems assurance engineers.

End Notes

In my quest for articles involving dependability, I have observed that modern society is becoming more dependent on technology, more dependent on people doing the right things, and, mostly, more dependent on identifying potential problems at design stage. Any systems affecting critical infrastructure are of particular importance.

Many industrial sectors need good dependability cases (and peer reviews thereof) more than ever, as incidents resulting in system downtime increasingly have major impact on business continuity and/or society. Next time you spot potential for a critical system or business going 'out of action', remember the OP bit.

Sometimes there's also safety implications – as has always been the case – but, without a formal dependability management regime/framework in place alongside the safety and human factors i.e. an integrated RAMS and HFI approach, together with management or duty holders who genuinely support such efforts, reliability and safety still have the means to tread a delicate, or even difficult, balance.

References

[1] https://www.theguardian.com/world/2023/jun/21/titanic-submersible-documents-reveal-multiple-concerns-raised-over-safety-of-vessel, accessed Dec 2023.

[2] https://www.news.com.au/national/nsw-act/news/single-sick-staff-member-sends-sydney-trains-network-into-chaos/news-story/0c28a932125aa79e2795f028155d5e7a, accessed Dec 2023

[3] https://www.bbc.co.uk/news/uk-66685349, accessed Dec 2023.

[4] https://www.bbc.co.uk/news/world-australia-67340901, accessed Dec 2023.

Image attribution:
Titan controller: https://en.m.wikipedia.org/wiki/File:Logitech_F710.jpg CC BY-SA 3.0
Titan endcap removal: https://en.wikipedia.org/wiki/Titan_submersible_implosion#/media/File:231001-G-G0000-1001.jpg public domain
Sydney departures board: https://en.m.wikipedia.org/wiki/File:Central_Station_indicator_board.jpg CC BY-SA 3.0
Heathrow tower: https://en.m.wikipedia.org/wiki/File:Heathrow_Traffic_Control_Tower_-_geograph.org.uk_-_144749.jpg, CC BY-SA 2.0
Triple zero: Photo 65600746 © Ymgerman | Dreamstime.com

Wendy Owen

Wendy has 35 years' experience in systems assurance across a wide range of highly regulated industries covering energy, defence, transport and process sectors. Her career has largely been in engineering consultancy, also with time in product development, research and regulatory. She is a Fellow of WES and the Safety & Reliability Society, and within the SCSC she is a task lead in SiSWG, a member of the Steering Group and lead for the Diversity Equity and Inclusion (DEI) initiative. She is currently freelance.

The author retains copyright of this article.


Safety Systems
2023 Compendium
Safety Systems
CRISIS MANAGEMENT
MINDSET THEORY
Managing the Safety & Security risks of Autonomous Vehicles
RADISH SEEDS
A new tool for managing data safety risks
For everyone working in Systems Safety
SCSC
thescsc.org
The Safety-Critical Systems Club Newsletter
Safety Systems
PORTRAITS OF WOMEN
Celebrating the women in System Safety
WELL MET!
HEAVY-DUTY SAFETY
Exploring Heavy-Duty Mobile Machinery
For everyone working in Systems Safety
SCSC
thescsc.org
The Safety-Critical Systems Club Newsletter
Safety Systems
STANDARD GAUGE
Recent developments in railway standards
IT'S A ZOO OUT THERE!
Developments in categorising data risk
ENCHANTRESS OF NUMBERS
Reflections on Ada Lovelace Day
For everyone working in Systems Safety
SCSC
thescsc.org
For everyone working in Systems Safety
SCSC
thescsc.org

scsc.uk/r192

Seminar: How to write Compelling Safety Arguments



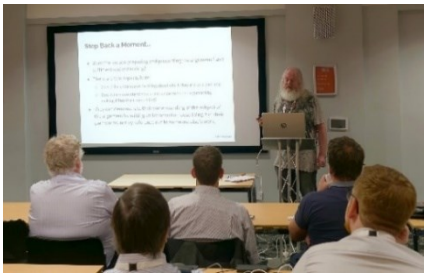
The "How to write Compelling Safety Arguments" seminar was held on the 28th September 2023 at the Wellcome Collection, London. The seminar looked at what makes a compelling safety argument, and how to use narrative techniques to improve the flow and 'story' of the argument. Christine Harper provides a summary of the event.

The seminar was hosted by Mike Parsons and the day consisted of the following four speakers who took different, but thought-provoking, approaches to "compelling" and "argument":

- John Spriggs (Independent Author and Presenter) "Sir, I have found you an argument…"
- Tim Rowe (Consultant) "It was the safest of systems; it was the most dangerous of systems. Textual analysis for better safety cases"
- Carmen Carlen (Edge Case Research (ECR)) "An Overview of Best Practices for Writing Compelling Safety Arguments"
- Emma Darwin (author) "Creating Engaging Safety Narratives – Narrative and Story"

The day also included a workshop led by Emma Darwin, "Workshop Introduction: Develop Your Own Safety Stories", as a short group exercise in order to try out some of the suggested approaches for improvement and included plenty of opportunities to network.

"Sir, I have found you an argument, but I am not obliged to find you understanding…"



John's title is based on a quote by James Boswell in his biography of his friend, Samuel Johnson, and also excellently summed up John's task.

The aim of a Safety Argument (SA) is to persuade the reader that the subject of the SA is safe. But readers will read a SA for different reasons, are from different backgrounds and may be satisfied by different amounts of evidence before they can accept the SA. This means that one cap does not fit all, and what may persuade one reader will be insufficient to persuade another. Safety Engineers rarely have the luxury of time to write a SA for each type of reader, so one SA must nevertheless suffice. Structuring the SA as:

- Easy to find points, or layers of sub-SA limited to just 1-2 layers
- Clearly bounding the SA

so that it is clear whether the SA supports ongoing use, or a move into service for example, is one way to get around the problem in one document.

However, no matter the structure, John noted that there is no place for wishful thinking in a SA, and ensuring that the supporting evidence does not interrupt the flow of the SA, is not contradictory to other supporting evidence and that it does not undermine the SA, are also key.

John stated that, "the SA evidence is only as good as the weakest link" ie. not all evidence is equal, and that the presented evidence might well be a subset of the available evidence.

All obvious stuff but the examples were very useful, and the hints to reviewers invaluable! My favourite was, "if evidence x were missing does it diminish the force of the argument?" which should enable me to happily prune away the dead wood in any future safety justifications I'm required to write or review!

Textual Analysis for a Better Safety Case

Tim's view is that a SA may be clear, concise, evidential and complete, but very boring so they are often not read as they should be. His answer: a plot, or a quest, ie a narrative or in fact a story, as in fact anything is better than none. And people enjoy and remember a good story! He also suggested pairing with some techniques more commonly associated with English literature:

- Use more natural language (especially less passive verbs and more active verbs)
- Actively seek "plot holes"
- Don't lie or mislead
- Remove material not necessary to the narrative off to somewhere else, and
- Consider using the journalists' pyramid structure in structuring the SA to allow the reader to stop reading when they feel they have read enough, and are persuaded

My takeaway: what better structure for the Exec Summary than a story!

Overview of Best Practices for Writing Compelling SA

Carmen's presentation was not so much about compelling plots and emotive adjectives but a logically compelling SA. Carmen divided SA into 3 types:

- Process-based (we followed and comply with standards, and here's the certificate)
- Product-based (we satisfy the requirements in this context and here is the evidence so it's ok)
- Confidence-based (the argument with accompanying sufficient evidence establishes a direct claim that it's ok)

She also used standard UL4600 as a checklist to help establish whether a SA plausibly satisfies system safety goals, and whether those system safety goals are sound. Very thought-provoking!

However, Carmen continued to move swiftly on. She briefly mentioned ECR's analysis of over 250 GSN based SA patterns used across different industries to comply with different standards (a resource lots of us thought it might be useful to explore) as a basis for building their safety case framework – the Open Autonomy Safety Framework (OASF), but reiterated that the framework is only based on GSN patterns, so doesn't include other approaches such as Claim Argument Evidence (CAE) patterns.



She also skated gracefully across automated checking of SA models and use of Safety Performance Indicators (SPI) when verifying or validating an SA. A huge amount of ground to cover in a short space of time!

I particularly liked her safety argument strategy:

- 'Live it right': our organisation is trustworthy
- 'Built it right': we've mitigated the risks
- 'Operate it right': we operate and maintain it responsibly
- 'Communicate it right': we have achieved a positive trust balance, evidenced by the sub arguments; the system is safe to deploy

Probably the perfect generic SA?

Narrative and Story



Emma's presentation was a really fascinating writing lesson that applies not just to creative writing but also to academic and technical writing. She stated that 'Human beings are wired for stories'.

She presented a number of approaches that you could use by yourself, or as a group, to build up and write your SA.

She defined a narrative as "a causally related chain of events", in other words, not a series of snapshots, and likened the story as "the journey you make" (comprising Beginning, Middle and End) and the plot (or narrative) as "the route you take" (the events and why they matter). Together these formed the basis of her approach to writing, as proposed in the Diamond template:

- The Thesis (main point or reason the document is being written)
- The Context/Explain (the things you need to know to follow, or understand or be convinced of)
- The Evidence/The Spine ("units of persuasion" – each with a Beginning, Middle and End; which can be fractal ie., composed of smaller units of persuasion)
- The Analysis and Interpretation (the "so what?" points from the units of persuasion in the spine)
- The big So What? (summing up the significance for the overall topic or situation, and where to go next)

Couple this with Emma's hints for journalists (who, what, where, when, how and why), the "…therefore …. but" approach to linking the parts of the narrative and her drafting/reviewing process (be yourself and decide what needs to be said, then check, will the reader get it, and finally check whether the person you need to persuade not just buys it, but buys into it) and the whole lot is dynamite!

Write Your Own Safety Story!

The group exercise to, "write your own safety story" complete with encapsulation phrase was a ball, resulted in some amusing children's stories (perfect for our general reader, Joe Public?) and allowed us all to have fun and participate! Check out some of Emma's blogs [1] when you have time!

My main takeaways from the day – the more SA you have written and reviewed, the more you took away from this day, and metaphors and allegories in a structural narrative can have a place in making a dry SA more memorable! Not a bad thing at all if your readers are happy to join you in the fun!

References

[1] Emma Darwin, This Itch Of Writing: The Blog, https://www.emmadarwin.com/this-itch-of-writing, accessed, Nov 2023

I would like to thank Richard Trebilcock for his contribution in sharing his notes of the event.

Seminar: What Makes Good Evidence for a Safety Case



The "What Makes Good Evidence for a Safety Case" seminar was held on 30th Nov 2023 at the Wellcome Collection in London. This seminar examined the role of evidence in supporting a safety case and explored the differences between compelling and weak evidence. It also considered what could be done to fill the gaps if evidence is not available or of poor quality or if evidence is continually evolving. The SCSC Newsletter Editor provides a summary of the day.

Mike Parsons opened the seminar and chaired proceedings, introducing each speaker and facilitating questions from the audience after each talk. The talks covered a wide range of domains including Aviation, Automotive and Healthcare. The event concluded with a lively open group discussion on a range of topics.

Evidence for a Satellite Datacomms to Civil Aircraft System



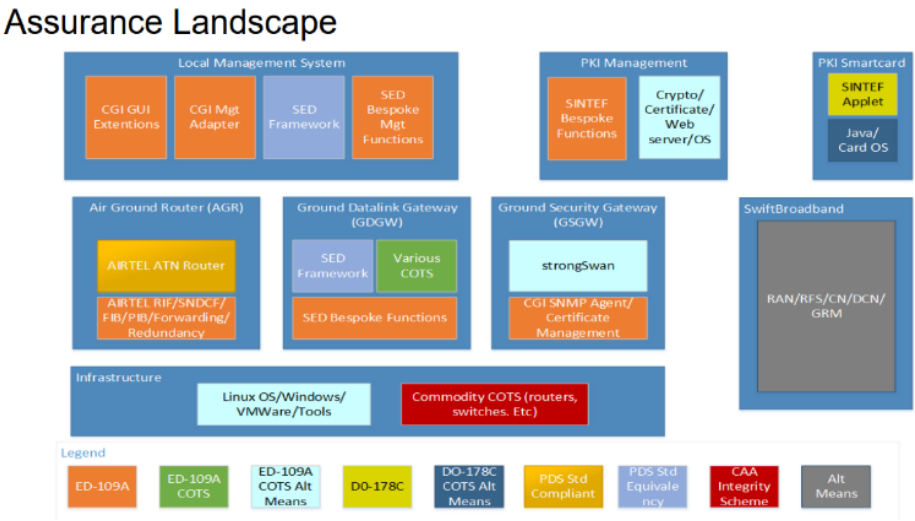
Paul Hampton from CGI gave the first talk of the day based around a system providing a data link for air traffic controller communications to commercial airline pilots. Paul discussed the risks and challenges of standards-based assurance when the system comprises subsystems of very different natures such as new, legacy, predeveloped software and software where alternative means of compliance arguments are required.

Paul first introduced the Iris system, which he has been working on for almost a decade and is now being deployed on live commercial flights. The system provides a SATCOM-based data link to supplement traditional VHF voice controller/pilot communications, which are seeing channels become increasingly congested. The use of data-based communication links also paves the way to more efficient route planning that will drive fuel saving costs and lead to reduced CO_2 emissions.

Paul provided the history of the project noting that when the project began, data link standards were not mature and did not cater well for SATCOM as a bearer for the data. The focus was therefore on the four assurance "pillars" of: safety, security, performance and interoperability and Paul showed how a structured approach to understanding the lifecycle of each of these assessment processes, and how they interacted, was key to developing an overall assurance argument for the entire system. Paul noted that the result was a very large and complex assurance argument and tools were required to help make the assurance case more accessible.

Paul then went on to explain how the regulation finally matured into requiring a 'Safety Support File' from the communication service provider that required evidence to demonstrate that: "the system would operate only as specified and would continue to do so in live service". Paul noted that producing such a case was akin to a traditional 'V' model and was well aligned with the earlier work that focused on explicitly breaking out each distinct phase of the development lifecycle from design through into service. Much of the evidence generated could then be allocated to different parts of the 'V' model to help support the overall assurance case.

Paul then went on to describe the many different assurance approaches adopted on the project. Although nominally, the system was to be developed to aviation standards, only the new software developments could be assured in this traditional way, as much of the system was legacy, COTS or based on previously developed software. Paul noted that nine different methods were used in the end to provide an equivalency to the level of assurance required by the standards.



The "assurance landscape" showing all the system components and which assurance approach (each with a different colour) was adopted for that component.

Paul concluded his talk by highlighting his thoughts on what makes good evidence:

- "Goodness" is not a property of evidence itself … but more on the narrative that accompanies that evidence
- Favour evidence of "least surprise" – evidence following a standard's objectives or known techniques will be more accessible to regulator/reviewer
- Adopt a structured approach to organising and presenting the evidence
  - Be clear on methodologies adopted and their interactions
  - Build evidence incrementally on smaller, manageable lower-level arguments
- "Less" can often be "more" – avoid temptation to focus on quantity, it can just overwhelm
- Be open about assurance deficits, but then argue why "that's ok" or provide plans to address

Addressing the Challenges of Evidence in Healthcare IT



In the next talk, Sean White from NHS England set the scene by saying healthcare was different from traditional safety as it considers the safety of poorly people rather than focussing on the existing safe state of, say, an aircraft.

He then discussed the Medical Device Regulation (MDR), which applies to products such as smart watches. The MDR uses a classification scheme that involves increasing rigour including independent

review as risk increases. He noted however, that MDR is for products and doesn't cover traditional clinical services Health IT.

Health IT systems are now covered by their own Data Coordination Board (DCB) standards that are now mandated: DCB0129 for Health IT manufacturers and DCB0160 for the deploying health organisation. These are based on ISO/IEC 14971 but include requirement for a Clinical Safety Officer (CSO) and a Safety Case Report.

Sean covered some of the challenges and which forms of evidence might help address those challenges.

Subjective Assessment of Risk

There is variance in risk schemes experience and corporate attitude to risk leading to subjective assessment of risk.

We would therefore seek to have more around explanation of the risk scheme used and rationale for the severity and scale of impact.

Qualitative determination of Risk

There is no quantitative assessment of risk for software-based products and processes are not prescribed.

Evidence should therefore be more around the justification: why the software development process will deliver the required integrity and engagement of the CSO in the process.

Reliance of "clinician in the loop"

There is a perception that the clinician can always intervene before things go wrong but we need to be confident that this is indeed the case.

Evidence should focus on engagement of the clinicians, their training and competency and defining pathway as 'work as done' to ensure there are credible opportunities for intervention.

Perceived Safety Significant

Consequences of failures are in general less severe and major "crash" events are rare. This leads to a culture of playing down risk.

More is required on the justification to ensure the clinical context is explained and there is close control on incidents so they can be properly managed.

Complex Architectures

Systems are becoming increasingly complex with cloud, mobile and variance in use cases. Development methodologies like agile are also adding to complexity.

Evidence should therefore focus on understanding system design, justifying 3rd party products and defining the development processes.

Safety Management Competence

Sean noted that there are issues with awareness and commitment to the standards and they are often retrospectively considered.

The focus should therefore be in training; ensuring top management understand their responsibilities and there is sufficient resourcing of, for example, an influential CSO.

Sean described the work being undertaken to de-risk the deficiencies:

- England is the first country to mandate a risk management process
- There is now specific guidance in a healthcare context for managing data safety risk
- There is a national portfolio of training addressing the importance of evidence
- There is a focus on proportional risk management; with limited resource the focus is on greatest risk
- The establishment of a CSO is important in providing domain knowledge, influence and direction, responsibility and accountability.
- Health organisations are encouraged to build DCB into contracts to ensure safety is built into the service

But he ended saying that inevitably it is a "leap of faith".

Sean concluded with the following observations:

- Problematic safety evidence is a symptom of other issues
- There is a need to establish confidence rather than absolute proof
- There is a need for more emphasis on pre-emptive risk management
- There is a need for more openness across organisational boundaries
- AI will compound existing challenges and introduce new ones

Safety Evidence: The Good, The Bad, The Ugly

Helen Wright from Leonardo said that not enough thought is, in general, given to the production of evidence in a safety case and there are issues with confirmation bias as highlighted by Nancy Leveson and the Nimrod Report.



She said we ideally want the case to be comprehensive, compelling and valid but they tend to be disjointed, difficult to navigate and end up trying to find large volume of evidence, a process which can end up pulling on threads that unravel the case. Helen thought that we need to raise the bar on what good safety evidence looks like.

Helen gave an example of safety activities undertaken on a civil aircraft programme that followed SAE ARP 4754/4761 processes in support of aircraft certification; this sits outside of a typical safety case regime. Combined safety/design generated extensive evidence although this was a lengthy process. She gave an example where safety analysis prompted a redesign of the fuel tank boundaries. She noted that the Aircraft Safety Synthesis was the closest thing to a safety case but without the argument. The same approach was implemented on the next programme designed for military operations, however in this example Helen highlighted some of the issues encountered when subsequently trying to adopt the safety case approach (as required by the Military Aviation Authority (MAA)) for release to

service with the UK MoD. The MAA requirement for a safety case and to be able to declare system risk reduced As Low As Reasonably Practicable (ALARP) and tolerable was substantially different to the evidence set required to support certification; this presented a number of challenges:

- The need to prepare a safety argument using available Certification evidence produced through a different process where appropriate and identify the potential evidence shortfalls in order to address them
- The additional need to produce a Hazard Log to support risk-based decisions at an appropriate level and employment of additional tools (such as Bow Tie Analysis) to aid Stakeholder engagement and facilitate communication to meet different regulatory requirements

This example was used to highlight the need to consider the regulatory environment and challenge the assumption that evidence used for one application might not necessarily be suitable in another.

Helen's final example was a safety-critical software project where assumptions about reuse of existing code were made during the bid phase such that only limited updates would be required to the safety analysis, the Safety Case Report (SCR) and the Safety Management Plan. This turned out not to be the case and substantial effort was required to unpick the traceability of requirements both from a safety analysis and a compliance with standards perspective as a result of seemingly insignificant changes in the expanded customer scope/requirements. The issues were largely resolved by identifying safety evidence requirements and detailed safety planning identifying the dependencies into the safety deliverables and feeding back into the system engineering lifecycle. The key learning point was that Safety Evidence Requirements instantiated into the project worked well.

Helen then concluded by discussing how to make things better and suggested:

- Establish confidence in the evidence from the start and identify any potential shortfalls
- Detailed safety planning as early as possible that is integrated with the system lifecycle and identifies dependencies (inputs to the safety analyses) that are generated by the wider team
- Challenge ambiguity in review, especially assumptions in the early stages of a project that influence the safety approach
- Need to get others on board by improving awareness of the evidence that will be required to support the safety argument – could use safety evidence requirements to make it explicit

Helen also suggested some other possible ideas to improve:

- Include confidence criteria in Goal Structuring Notation (GSN) and counter safety-arguments
- Eliminative argumentation and adopt a more mathematical approach
- It may be better to focus more on why it is not safe
- Improving clarity in communication of hazards with Stakeholders to ensure they have suitable and appropriate information on which to make their decisions

Autonomous Vehicles and Safety Case Evidence



Rebecca Wilford and Norina Ratiu from Oxa discussed their work in developing an L4 autonomous vehicle. Norina discussed the safety lifecycle across a 'V' model from project definition through to Safety in-use monitoring.

Autonomous Vehicles (AV) take on some of driving tasks such as strategic (mission specification), manoeuvre (behaviour stay-in lane etc) and control (throttle brake steering). The safety scope of AVs therefore needs to cover behaviour competencies, malfunctions, human misuse and cyber threats.

Norina and Rebecca discussed their Positive risk balance (PRB) approach. MISRA 2019 have developed guidelines for Automotive Safety Arguments. This splits the argument into 4 claims: Means, Rationale, Satisfaction and Organisational Environment.

They then gave examples of supporting claim themes using GSN for each of these four areas. They then discussed evidence quality and gave examples of the different types they encounter:

- **Best**: completeness in documentation and configuration control with traceability and evidence of review
- **Average**: summaries, front pages only, partially controlled, in note form
- **Bad**: Verbally communicated only, second hand sources or implied

Keeping an Operational Safety Case Alive and Well



Andy Whalley from BAE Systems discussed air system safety cases in the Military domain.

He has responsibility for managing the safety cases of several aircraft and likened maintaining a safety case to raising young children where they need nurturing and eat a lot, that is, they require a lot of evidence. He also said it is also not something that is done and completed but needs to be maintained throughout operations.

He noted that end operational cases need to be thought about at the conceptual stage even though not much evidence will be there to start. There can also be two operational cases at one time, one for the airframer and one for the airline (including maintenance etc).

The Military Aviation Authority (MAA), the regulator in the military domain, likened safety to a bath – evidence heats up bath water to operational standard as evidence is added.

Andy covered 3 aspects of safety case design: risk, rules-conformance and confidence and discussed what evidence is important through each of the product lifecycle phase.

Design phase



Risk

- Understand and ensure the context is right; in particular looking for CONOPS, to explain the Who/What/Where/When
- Ensure design has safety targets, eg. quantitative loss targets for the given context

Rules/Conformance

- Be clear on which regulatory set are being worked to (there are about 300 in military aviation and be able to argue which ones are applicable and which have been disregarded)

Confidence

- Need to ensure there are competent contracted organisations
- Staff in the organisation will need to be SQEP and there should be accountable stakeholders
- Check there are clear plans, principles and processes in place

Build/Test Phase



The evidence 'bathtub' will now get slightly warmer as the project moves into the Build/Test Phase and there will now be test plans.

Risk

- System safety assessment can now be conducted to assess the design and determine the residual risk
- Trials risk assessments can also be conducted
- Assumptions/dependencies are important and often a forgotten aspect. Need to look for evidence that assumptions/dependencies have been accepted across stakeholders.

Rules/Conformance

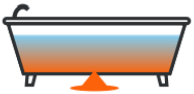
- Understand and handle non-compliances
- Compare the Design versus Build and versus Maintenance Standards to ensure they are consistent

Confidence

- Determine if targets have been met through post-test reports and risk-to-life review (assess residual risk)
- Undertake 1st 2nd 3rd party assurance, employ Independent Safety Assessors (ISA) and conduct regular safety/risk to life review

Andy provided further details on how to make the risk to life review through the building of a dynamic operational safety case based on a pyramid structure diagram. This allowed changes in the environment to be quickly assessed for impact on the safety of life assessment.

Entry-into-service and beyond



Once the system enters service the "bathtub" is now full of warm water, but it needs to be maintained to ensure the water does not get cold or leaks away.

Risk

- Ensure changes in configuration/environment, safety enhancing technologies/techniques, new technologies are managed and reflected in the safety case
- Ensure issues (the 'bath leaks') are managed (eg. operator or maintenance errors, accidents etc.) and review wider fleet occurrences

Rules/Conformance

- Monitor and assess the impact of regulatory change and conduct usage validation reviews

Confidence

- As with the previous phase: undertake 1st 2nd 3rd party assurance, ISA and regular safety/risk to life reviews.

Andy then looked at some of the future challenges:

- If AI has been used to generate evidence, how do we trust it? Similarly, what if it is used in the design, production and even in the safety case generation?
- Digital twins have also become more prominent in recent times. Could these be used for synthetic training of products? Virtual hazard analysis (interrogating a digital model) is already being done.

Andy concluded by noting that:

- Evidence matures/evolves through-life
- Understanding operational context is key for operational safety case
- Evidence can grow stale, which can then lead to an erroneous safety case

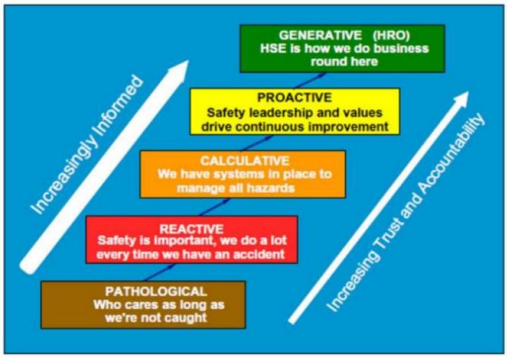
Confessions of an Auditor



Sarah Tovey from the MOD, described her roles in the independent technical services team and aimed to illustrate trends that were not domain specific.

The Journey

Sarah referred to the Parker-Hudson model, which is a model for safety culture in an organisation running from Pathological through to Generative.



Sarah thought the organisation was probably more reactive initially than "generative" and noted there were many issues such as not being consistent across domains, there were known problems and change programme upon change were disruptive.

3-5 years ago, the Safety and Environmental Management System (SEMS) documentation was being updated but not evolving quite as fast as they would have liked. 5 years ago, an in-house assurance team was developed that provided better oversight and robust evidence of trends, however there were still issues and new procurement models also complicated matters.

Where they are now

Sarah said they have now introduced consistent assurance across domains and internal delivery specialists established but there are still issues, known problems continue – and there is still change upon change.

Sarah described the Defence Safety Management System "wheel" from JSP185 that provides an SMS structure that is consistent across domains and then discussed trends and themes across some of these safety management areas.

Trends and themes

- Shelfware, lack of linkage of critical project milestones, quality variations, repetition of policy, resources, continuous improvement opportunities missed.
- Legislative compliance: deficiencies with demonstrating compliance and risk of deficiencies. Training systems and Compliance with government supplied equipment poorly addressed.
- Environmental management missing disposal plans.
- Different tools: risk to life v equipment contribution to risk of life. Issues with putting an air platform on a ship – are both parties evidence there?
- Underpinning evidence – when is enough enough? Some domains better than others.
- Personal competence resources and training: lack of SQEP, ongoing cycle of routine maintenance overlooked. Lack of succession planning. "niche" specialisms.
- Assumptions management; unknown risk as assumption that is unproven or not mitigated.
- Performance management information and reporting: what to report, trend management and learning from experience.
- Assurance: lack of resource and oversight.

Where they want to be

Sarah said that they want to be on the edge of calculative/proactive but noted this will take time but will require:

- Effective in-house and cross-domain assurance
- Embedded processes: internal delivery, change programme, psychological (people empowered to speak up)
- Known problems addressed

Takeaways and challenges

Sarah concluded with some final takeaways:

- It is an ongoing journey; cultural changes take time to embed
- Sharing of knowledge and consistency between projects could be improved; often see different teams making the same mistakes, which could have been prevented if there was better sharing

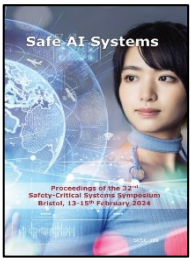
Group Discussion

The day ended with a group discussion to discuss what could be done to improve best practice. The following ideas were discussed:

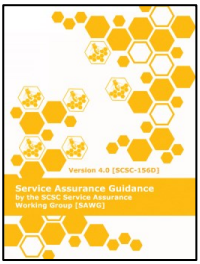
- Expiry dates for evidence
- Standardising the weighting of evidence
- Evidence can be a byproduct but cost benefit may be difficult to evaluate
- Can evidence be quantified in isolation without the narrative; perhaps needs to be about relevance?
- High building safety new to systems safety – currently more compliance based – but big changes have occurred with every building over 80m needing a safety case and other important changes around how we engage with the residents.
- Impact of AI how are these going to be handled in the future. All depends on how it is used. May end up just being reviewers of AI outputs rather than authors.
- Should we do a hazard analysis ourselves of where AI can be used safely?
- May need a separate safety checker of AI outputs. Should it be focused on just showing where there are holes in a safety case?
- Safety culture is more important than we think. Culture can destroy the evidence base if not appropriate.
- Who is the readership of the safety case? Agreed that it is best if crafted as a narrative and story to bring the reader along.
- Is good safety evidence something you can trace back and understand its traceability to requirements?
- Should we be able to learn from rare events too?
- How does an author know that evidence is good? Should it be a group effort?
- Who "owns" the safety case? Authors can't be expected to know everything is correct in all aspects. Needs to be a collective – seen as a safety team activity

Mike wrapped up the session and concluded by saying the SCSC will be launching an AI Working Group.

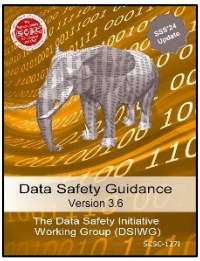
Recent Safety Publications



Proceedings of the 32nd Safety-Critical Systems Symposium.
scsc.uk/scsc-188



Service Assurance Guidance Version 4.0. Jan 2024
scsc.uk/scsc-156D



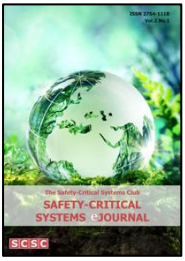
Guidance on the management of data safety risks (V3.6)
scsc.uk/scsc-127I



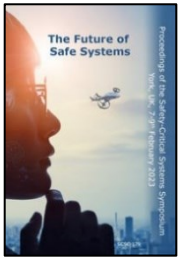
Safety-Critical Systems eJournal vol.3 no.1
Winter Issue 2023
scsc.uk/scsc-191



Safety Culture Working Group Position Paper
Edition 1. Dec 2023
scsc.uk/scsc-189



Safety-Critical Systems eJournal vol.2 no.2
Summer Issue 2023
scsc.uk/scsc-186



Proceedings of the 31st Safety-Critical Systems Symposium.
scsc.uk/scsc-179



Catastrophe and Systemic Change. Learning from the Grenfell Tower Fire and Other Disasters.
https://www.amazon.co.uk/dp/1913019292



Guidance on the Safety Assurance of autonomous systems in Complex Environments.
http://www.york.ac.uk/assuring-autonomy/guidance/sace/

60 Seconds with … ChatGPT



To tie-in with this edition's AI theme, I conducted a special 60-second interview with ChatGPT, the popular chatbot developed by OpenAI. I asked it to imagine it was a system safety engineer, and asked it the usual interview questions[4].

What first attracted you to working in the field of System Safety?

From the outset, I was drawn to system safety by its pivotal role in mitigating risks and ensuring the dependability of intricate systems. The prospect of contributing to accident prevention across diverse industries, from aviation to healthcare, fascinated me. As an engineer, the challenge of systematically identifying potential hazards and implementing preventative measures aligned with my problem-solving inclination.

The dynamic nature of the field, where each project demands a unique approach to risk management, offered a stimulating and intellectually rewarding environment. Moreover, the awareness that my work directly impacts user safety and public well-being added a profound sense of purpose to my professional journey. The blend of technical expertise, continuous learning, and the tangible impact on safety outcomes has made system safety an engaging and fulfilling career choice for me.

What aspect of your career are you most proud of?

I take immense pride in successfully spearheading the integration of a comprehensive safety program for a critical system. Leading a diverse team, we meticulously identified potential risks, strengthened vulnerabilities, and implemented robust safety protocols. Witnessing the system operate incident-free and knowing that our efforts directly contributed to user safety has been incredibly rewarding.

Additionally, I am proud of my commitment to continuous improvement, staying at the forefront of evolving safety standards and technologies. This dedication not only enhances the safety measures of specific projects but also contributes to the overall advancement of safety practices within our organization. These accomplishments highlight my passion for ensuring the reliability and security of complex systems, making a lasting impact on both project outcomes and the broader field of system safety.

[4] I also explored how ChatGPT might realise itself as a physical entity by asking it to suggest AI image creation prompts. It determined it should reflect: "*processing vast amounts of information and generating coherent responses, symbolising interconnectedness, intelligence, and adaptability, mirroring the qualities associated with advanced artificial intelligence.*" and the results are shown top left using these words as a prompt to the Midjourney AI art generating application.

Tell us one interesting unusual fact about yourself!

While my professional life is dedicated to ensuring the safety and reliability of complex systems, one interesting and somewhat unusual fact about me is my passion for extreme sports. Outside the realm of risk analysis and safety engineering, I thrive on adrenaline-fueled activities like skydiving and rock climbing. Engaging in these high-risk endeavours provides me with a unique perspective on the importance of safety measures not only in the professional domain but also in personal pursuits.

It's a thrilling juxtaposition – meticulously assessing and mitigating risks in my career while willingly embracing controlled risk and adventure in my personal life. This duality not only keeps me grounded but also underscores the significance of a balanced approach to risk, whether in the calculated design of systems or the exhilaration of pushing personal boundaries.

What advice would you give to yourself age 12?

Reflecting on my journey as a system safety engineer, I would advise my 12-year-old self to embrace curiosity and never shy away from exploring diverse interests. It's crucial to cultivate a passion for problem-solving and critical thinking early on, as these skills become the foundation of a successful career in safety engineering.

Don't be afraid to take on challenges and learn from failures—they are invaluable lessons. Foster resilience, adaptability, and a love for continuous learning, as technology and safety standards evolve rapidly. Surround yourself with mentors who can guide and inspire you. Most importantly, recognize the importance of collaboration and effective communication in any field. The ability to work harmoniously within a team is as vital as technical expertise. So, young self, stay curious, be bold in your pursuits, and never underestimate the power of a well-rounded and adaptable skill set.

What future changes would you like to see in the field of System Safety?

Looking ahead, I envision several positive changes in the field of system safety. Firstly, a more seamless integration of advanced technologies like artificial intelligence and machine learning into safety analysis processes would enhance our ability to predict and prevent potential risks.

Additionally, fostering a culture of interdisciplinary collaboration could lead to more holistic safety approaches, considering not only technical aspects but also human factors and organizational influences. I hope to see greater standardization and global harmonization of safety practices, facilitating smoother cooperation across industries and international borders. Embracing a proactive rather than reactive mindset, with an increased emphasis on pre-emptive risk identification and mitigation, is essential for the future. Lastly, continued efforts in education and professional development can ensure that the next generation of system safety engineers is well-equipped to navigate evolving challenges in this dynamic and critical field.

What's your most favourite quote or motto?

"One of my favourite mottos is 'Prevention is the best intervention.' In the realm of system safety, this encapsulates the essence of our work. By identifying and addressing potential risks before they materialize into issues, we not only ensure the reliability of complex systems but also prioritize the well-being and safety of users. It underscores the proactive nature of safety engineering, emphasizing the importance of thorough risk analysis and preventative measures. This motto serves as a constant reminder of the critical role we play in averting potential accidents and upholding the highest standards of safety in diverse industries.

If you could learn to do anything, what would it be?

As a system safety engineer, if I could learn anything beyond my current expertise, I would choose to deepen my understanding of human factors and psychology. Enhancing my knowledge in these areas would enable me to better integrate human-centred design principles into safety engineering processes. Understanding how individuals interact with complex systems, assessing cognitive biases, and improving user interfaces could significantly contribute to the overall safety and usability of the systems I work on. This interdisciplinary approach would allow me to address not only technical aspects but also human behaviours and decision-making, ultimately creating more effective and user-friendly safety solutions. It aligns with the belief that a comprehensive understanding of both technical and human factors is crucial for developing truly robust and reliable systems.

If you could be any fictional character, who would you choose?

As a system safety engineer, if I were to choose a fictional character, I would identify with Mr. Spock from Star Trek. His logical and analytical approach to problem-solving resonates with the systematic and risk-oriented mindset required in system safety engineering. Like Mr. Spock, I aim to make decisions based on a thorough analysis of facts and probabilities, prioritizing the safety and well-being of complex systems and their users. Additionally, his commitment to balancing logic with a sense of duty and ethical considerations mirrors the need for a well-rounded approach in the field of system safety. While fictional, Mr. Spock's character embodies qualities that align with the challenges and responsibilities inherent in my profession.

Artificial Intelligence - a positive revolution or are we all doomed!?

As a system safety engineer, I see Artificial Intelligence (AI) as a positive revolution with tremendous potential for enhancing safety and efficiency. While there are concerns about AI, responsible implementation can significantly improve various industries. In safety engineering, AI contributes to predictive analytics, identifying potential risks, and automating complex tasks. Properly designed AI systems can augment human capabilities, leading to more robust safety measures and quicker response times. However, cautious and ethical development is crucial to mitigate potential risks. As a community, we must establish robust guidelines, emphasizing transparency, accountability, and bias mitigation in AI systems. The future depends on our ability to harness AI's power responsibly, ensuring it aligns with safety principles and ethical standards. With thoughtful implementation, AI can be a transformative force, positively impacting system safety and numerous other aspects of our lives.

Connect

The Newsletter and eJournal

Do you have a topic you'd like to share with the systems safety community? Perhaps an interesting area of research or project work you've been involved in, some new developments you'd like to share, or perhaps you would simply like to express your views and opinions of current issues and events. There are now two publishing vehicles for content – shorter, more informal content, can be published in the Newsletter with longer, more technical peer-reviewed material more suitable for the eJournal. If you are interested in submitting content, then get in touch with Paul Hampton for Newsletter articles: paul.hampton@scsc.uk or John Spriggs for eJournal papers: john.spriggs@scsc.uk

The SCSC Website

Visit the Club's website thescsc.org for more details of the Safety-Critical Systems Club including past newsletters, details of how to get involved in working groups and joining information for the various forthcoming events.

Facebook

Follow the Safety-Critical Systems Club on its very own Facebook page.

www.facebook.com/SafetyClubUK

X/Twitter

Follow the Safety-Critical Systems Club's X/Twitter feed for brief updates on the club and events: @SafetyClubUK

LinkedIn

You can find the club on LinkedIn. Search for the Safety-Critical Systems Club or use the following link:

www.linkedin.com/groups/3752227

Advertising

Do you have a product, service or event you would like to advertise in the Newsletter? The SCSC Newsletter can reach out to 1,000's of individuals involved in Systems Safety and so is the perfect medium for engaging with the community. For prices and further details, please get in touch with the Newsletter Editor.

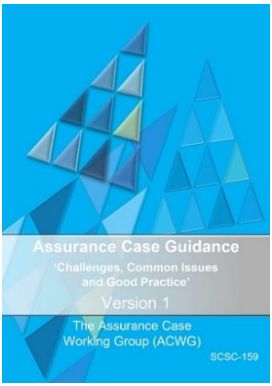
SCSC Working Groups

The Safety-Critical Systems Club is committed to supporting the activities of working groups for areas of special interest to club members. The purpose of these groups is to share industry best practice, establish suitable work and research programmes, develop industry guidance documents and influence the development of standards.

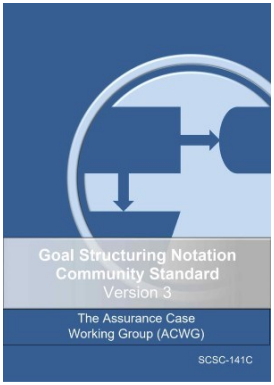
Assurance Cases

The Assurance Cases Working Group (ACWG) has been established to provide guidance on all aspects of assurance cases including construction, review and maintenance. The ACWG will:

- Be broader than safety, and will address interaction and conflict between related topics
- Address aspects such as proportionality, rationale behind the guidance, focus on risk, confidence and conformance
- Consider the role of the counter-argument and evidence and the treatment of potential bias in arguments



In Aug 2021, the group published v1.0 of the Assurance Case Guidance: scsc.uk/scsc-159



One of the working group's activities is the maintenance of the Goal Structuring Notation (GSN) Community standard.

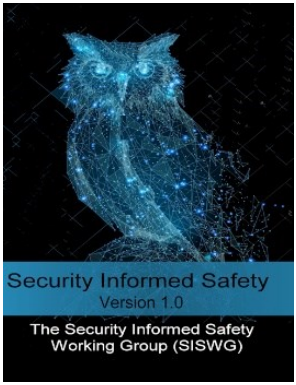
See scsc.uk/gsn for further details.

In May 2021, the group published v3.0 of the standard: scsc.uk/scsc-141C

Lead **Phil Williams** phil.williams@scsc.uk

SCSC Working Groups

Security Informed Safety



The Security Informed Safety Working Group (SISWG) aims to capture cross-domain best practice to help engineers find the 'wood through the trees' with all the different security standards, their implication and integration with safety design principles to aid the design and protection of secure safety-critical systems and systems with a safety implication.

The working group aims to produce clear and current guidance on methods to design and protect safety-related and safety-critical systems in a way that reflects prevailing and emerging best practice.

The guidance will allow safety, security and other stakeholders to navigate the different security standards, understand their applicability and their integration with safety principles, and ultimately aid the design and protection of secure safety-related and safety-critical systems.

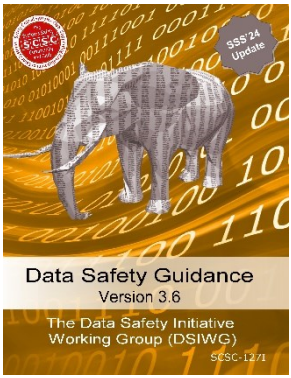
Lead **Stephen Bull** stephen.bull@scsc.uk

Data Safety Initiative

Data in safety-related systems is not sufficiently addressed in current safety management practices and standards.

It is acknowledged that data has been a contributing factor in several incidents and accidents to date, including events related to the handling of Covid-19 data. There are clear business and societal benefits, in terms of reduced harm, reduced commercial liabilities and improved business efficiencies, in investigating and addressing outstanding challenges related to safety of data.

The Data Safety Initiative Working Group (DSIWG) aims to have clear guidance on how data (as distinct from the software and hardware) should be managed in a safety-related context, which will reflect emerging best practice.

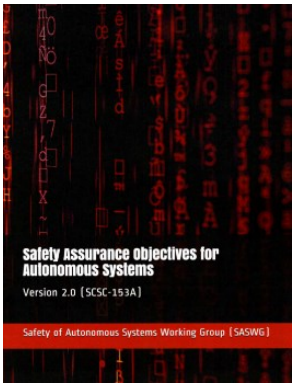


An update to the guidance (v3.6) was published in Jan 2024: scsc.uk/scsc-127I

Lead **Mike Parsons** mike.parsons@scsc.uk

SCSC Working Groups

Safety of Autonomous Systems



The specific safety challenges of autonomous systems and the technologies that enable autonomy are not adequately addressed by current safety management practices and standards.

It is clear that autonomous systems can introduce many new paths to accidents, and that autonomous system technologies may not be practical to analyse adequately using accepted current practice. Whilst there are differences in detail, and standards, between domains many of the underlying challenges appear similar and it is likely that common approaches to core problems will prove possible.

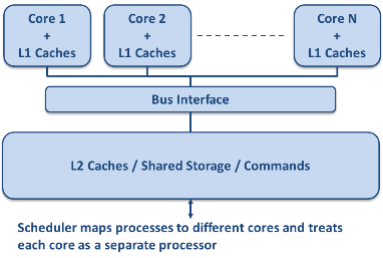
The Safety of Autonomous Systems Working Group (SASWG) aims to produce clear guidance on how autonomous systems and autonomy technologies should be managed in a safety-related context, in a way that reflects emerging best practice.

The group published v3 of its guidance Safety Assurance Objectives for Autonomous Systems, in Jan 2022 scsc.uk/scsc-153B

Lead Philippa Ryan pmrc@adelard.com

Multi- and Manycore Safety

It is becoming harder and harder to source single-core devices and there is a growing need for increased processing capability with a smaller physical footprint in all applications. Devices with multiple cores can perform many processes at once, meaning it is difficult to establish (with sufficient evidence) whether or not these processes can be relied upon for safety-related purposes.



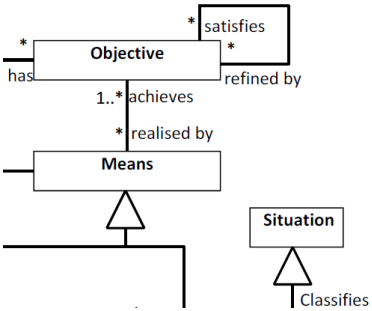
Parallel processes need to access the same shared resources, including memory, cache and external interfaces, so they may contend for the same resources. Resource contention is a source of interference which can prevent or disrupt completion of the processes, meaning it is difficult to know with a defined uncertainty the maximum time each process will take to complete (Worst Case Execution Time, WCET) or whether the data stored in shared memory has been altered by other processes.

The Multi- and Manycore Safety Working Group (MCWG) has been established to explore the future ways of assuring the safety of multi- and manycore implementations.

Lead Lee Jacques Lee.Jacques@leonardocompany.com

SCSC Working Groups

Ontology



The Ontology Working Group (OWG) develops ontologies that will form the basis of SCSC guidance, as well as having wider industrial and academic applications.

The OWG is currently working on the definition of an ontology of risk for application in guidance for risk-based decision making – notably safety and security – and for which ISO 31000 Risk Management principles are to be applied.

The Data Safety Working Group (DSIWG) developed the core aspects of the Risk Ontology, which has been migrated to this working group. The Risk Ontology will form the upper ontology to the Data Safety Ontology that the DSIWG will continue to develop.

Lead Dave Banham ontology@scsc.uk

Service Assurance

Risks presented by safety-related services are rarely explicitly recognised or addressed in current safety management practices, guidelines and standards. It is likely that service (as distinct from system) failures have led to safety incidents and accidents, but this has not always been recognised. The Service Assurance Working Group (SAWG) has been set up to produce clear and practical guidance on how services should be managed in a safety-related context, to reflect emerging best practice.



The group published v4.0 of the guidance in Jan 2024: scsc.uk/scsc-156D

Lead Kevin King kevin.king@baesystems.com

Safe AI

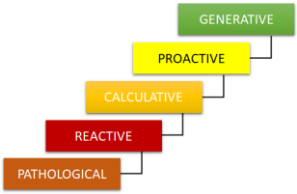
The SCSC Safe AI Working Group (SAIWG) will aim to capture cross-domain best practice and guidance on key topics within the design, evaluation, assurance, and approval of safety systems that use or are developed using AI, bringing together emerging standards and key results from the incredible amount of research being conducted into AI safety.

The working group will kick-off at SSS'24 and will be led by Alan Simpson. From then the SAIWG will conduct regular meetings, workshops, and publications to share knowledge and experience on various topics related to AI and safety systems, such as coordination of safety with other disciplines, evaluation of risk, and mapping of terminology and language.

Joining instructions will be announced at SSS'24.

SCSC Working Groups

SCSC Safety Culture



The Safety Culture Working Group (SCWG) has been established to provide guidance on creating and maintaining an effective safety culture. The group seeks to improve safety culture in safety-critical organisations focussed on product and functional safety, by sharing examples and latest approaches collated from real-life case studies.

Meetings provide an opportunity to discuss any particular aspects attendees are interested in taking forward, and to help set future directions for the group.

In Dec 2023, the group published a position paper for assessing and managing safety culture. https://scsc.uk/r189:1

Lead Michael Wright michael.wright@greenstreet.co.uk

Systems Approach to Safety of the Environment



The Systems Approach to Safety of the Environment Working Group (SASEWG) is a new group intending to apply Systems Safety practices to systems that are embedded within the natural environment, while focussing on that environment.

The group aims to produce clear guidance on how engineered systems should be developed and managed throughout their entire lifecycle so as to preserve, protect and enhance the environment.

Since holding its inaugural meeting on 17th June 2022 the group is now well underway. Please get in touch with the working group lead if you would like to join or find out more about this group.

Lead James Inge james.inge@scsc.uk

Safer Complex Systems



The Safer Complex Systems Working Group (SCSWG) builds on the IET/RAE work already done in this area. It is recognised that the RAE work is ongoing and collaboration is encouraged. The group's mission is to produce practical guidance on developing, managing and assuring complex systems throughout their lifecycle (so as to achieve and justify their safety).

Contact the group lead if you would like to attend future meetings or find out more about the working group.

Acting Lead Mike Parsons mike.parsons@scsc.uk

SCSC Membership

The SCSC provides a range of services to the System Safety community including seminars, tutorials, leadership events, specialist topic working groups, the annual symposium and a comprehensive body of publications. Membership brings many valuable benefits such as free access to online events, the SCSC Newsletter and access to presentations and other resources from events.

Individual Membership

To become an individual member of the SCSC please register on the SCSC website using the icon at the top right of any page and select "Register". Complete and save your account registration and then verify your email address. Once registered and logged in click the link "why not join the SCSC..." inviting you to become a member at the top right of the page or select "Pay membership" from the icon.

Individual membership can be paid online using a credit/debit card through our secure payment partner Realex Global Payments or contact Alex King for other payment methods. For student or retired member rates please contact Alex King to get your account status changed.

Corporate Membership

Your company contact with the SCSC should arrange the membership and any renewals for your organisation. To join as a member covered by a corporate membership, register as per the instructions for an individual member and then contact Alex King to confirm your affiliation.

Renewing Membership

You should be notified by email when your membership is almost expired or shortly after it has expired. These notifications will contain a link to the online renewal page or you will be able to renew when logging onto the website through the 'click to renew' link.

Membership Fees

The following fees are applicable for new and renewing members:

- 1 year Individual Membership: £145
- 2 year Membership: 13% discount: £250
- 3 year Membership: 19% discount: £350
- 1 year SFI Membership: FREE for first year, £35 for years 2 & 3
- 1 year Membership, retired member rate: £35
- For Corporate Membership discounts contact Alex King

A one-month Publication Pass is also available for £15. This allows access to all SCSC website publications in a particular calendar month.

Contact Alex King using office@scsc.uk

The SCSC Steering Group


Tom Anderson
Honorary member


Stephen Bull
stephen.bull@scsc.uk


Dai Davis
Honorary member


Zoe Garstang
zoe.garstang@scsc.uk


Louise Harney
louise.harney@scsc.uk


Brian Jepson
brian.jepson@scsc.uk


Graham Jolliffe
Honorary member


Alex King
alex.king@scsc.uk


Wendy Owen
wendy.owen@scsc.uk


Davy Pissoort
davy.pissoort@scsc.uk


Roger Rivett
roger.rivett@scsc.uk


Emma Taylor
Honorary member


Sean White
sean.white@scsc.uk


Robin Bloomfield
Honorary member


Dewi Daniels
dewi.daniels@scsc.uk


Jane Fenn
jane.fenn@scsc.uk


Paul Hampton
paul.hampton@scsc.uk


James Inge
james.inge@scsc.uk


Nikita Johnson
nikita.johnson@scsc.uk


Tim Kelly
Honorary member


Mark Nicholson
mark.nicholson@scsc.uk


Mike Parsons
mike.parsons@scsc.uk


Felix Redmill
Honorary member


John Spriggs
john.spriggs@scsc.uk


Phil Williams
Honorary member

Club Positions

The current and previous (marked in italics) holders of club positions are as follows:

Managing Director

Mike Parsons 2019-

Tim Kelly 2016-2019

Tom Anderson 1991-2016

Programme & Events Coordinator

Mike Parsons 2014-

Chris Dale 2008-2014

Felix Redmill 1991-2008

Newsletter Editor

Paul Hampton 2019-

Katrina Attwood 2016-2019

Felix Redmill 1991-2016

eJournal Editor

John Spriggs 2021-

Diversity, Equity and Inclusion (DE&I) Lead

Wendy Owen 2023-

Steering Group Chair

Roger Rivett 2019-

Graham Jolliffe 2014-2019

Brian Jepson 2007-2014

Bob Malcolm 1991-2007

Manager

Alex King 2019-

Honorary Solicitor

Dai Davis 2022-

University of York Coordinator

Mark Nicholson 2019-

Website Editor

Brian Jepson 2004-

Administrator

Alex King 2016-

Joan Atkinson 1991-2016

Safety Futures Initiative Leads

Zoe Garstang 2019-
Nikita Johnson 2023-

Nikita Johnson 2019-2021

Calendar

January

M	T	W	T	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

February

M	T	W	T	F	S	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29			

March

M	T	W	T	F	S	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

April

M	T	W	T	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

May

M	T	W	T	F	S	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

June

M	T	W	T	F	S	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

July

M	T	W	T	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

August

M	T	W	T	F	S	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

September

M	T	W	T	F	S	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

October

M	T	W	T	F	S	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

November

M	T	W	T	F	S	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

December

M	T	W	T	F	S	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Events Diary



13-15 Feb 2024
SCSC Symposium

Safety Critical Systems Symposium SSS'24

Bristol, UK

scsc.uk/e1007

8-11 April 2024
Conference

19th European Dependable Computing Conference (EDCC'24)

Leuven, Belgium

edcc2024.esat.kuleuven.be

20 April 2024
SCSC Tech Trip

Visit and Guided Tour of the Shuttleworth Collection

Biggleswade, Bedfordshire

scsc.uk/e1101

25 April 2024
SCSC Seminar

Seminar: Frameworks for Safe AI Systems

London, UK

scsc.uk/e1081

13-14 May 2024 (Part 1)
Training Course

SCS: IET Safety Critical Systems training course

Online

safetysystems.theiet.org

20-21 May 2024 (Part 2 & 3)
Training Course

SCS: IET Safety Critical Systems training course

Online

safetysystems.theiet.org

4-7 June 2024
Symposium

Safe Process

Ferrara, Italy.

safeprocess2024.eu

12-13 June 2024
Conference

Safety of Industrial Automated Systems (SIAS'24)

Tampere, Finland.

automaatioseura.fi/sias2024

23-27 June 2024
Conference

34th European Safety and Reliability Conference (ESREL'24)

Cracow, Poland.

esrel2024.com

24-27 June 2024
Conference

The 54th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'24)

Brisbane, Australia.

dsn2024uq.github.io

27 June 2024
SCSC Seminar

Seminar: Developing Safe AI Systems

London, UK

scsc.uk/e1100

17-20 Sept 2024
Conference

43rd International Conference on Computer Safety, Reliability and Security (SafeComp)

Florence, Italy.

safecomp2024.unifi.it


thescsc.org/membership
Safety Critical Systems Club Supporting Development of Safer Systems
The Systems Safety Community and Club
SCSC