[dstl]
Delivering
Mission Success

AI and the Agnogenic Test Engineer

Dr Al Banks

Principal Advisor Autonomy and AI Safety Engineering

Ministry of Defence

21/10/2026 / © Crown copyright 2026 Dstl

DSTL/PUB181071

OFFICIAL

Test Engineering – Not Me Chief, I'm Airframes…

Testing: involves ensuring that requirements have been fully met by using appropriate tools and techniques to verify that a product or service works.*

- Do I do testing?
  - Absolutely!
- Do I need to worry about this?
  - Absolutely!
- Is this presentation going to tell me everything I need to know?
  - Absolutely not!



AI Image: GPT-5

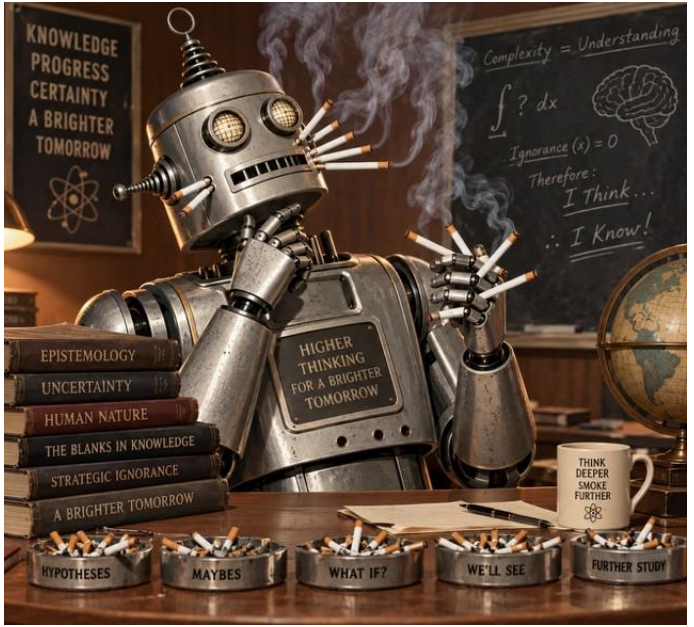
*His Majesty's Government: https://ddat-capability-framework.service.gov.uk/skills#test-engineering

OFFICIAL

Agnotology – What You've Been Dying To Know…

Agnotology: the study of the creation, maintenance, and dissemination of ignorance, doubt, or uncertainty

- The concept emerged during research into the tobacco industry's efforts to hide the risk of smoking*
- Key insight: ignorance can be *natural or created*
- Concept has been extended to domains such as:
  - Climate change denial
  - Strategic misinformation
  - Institutional secrecy



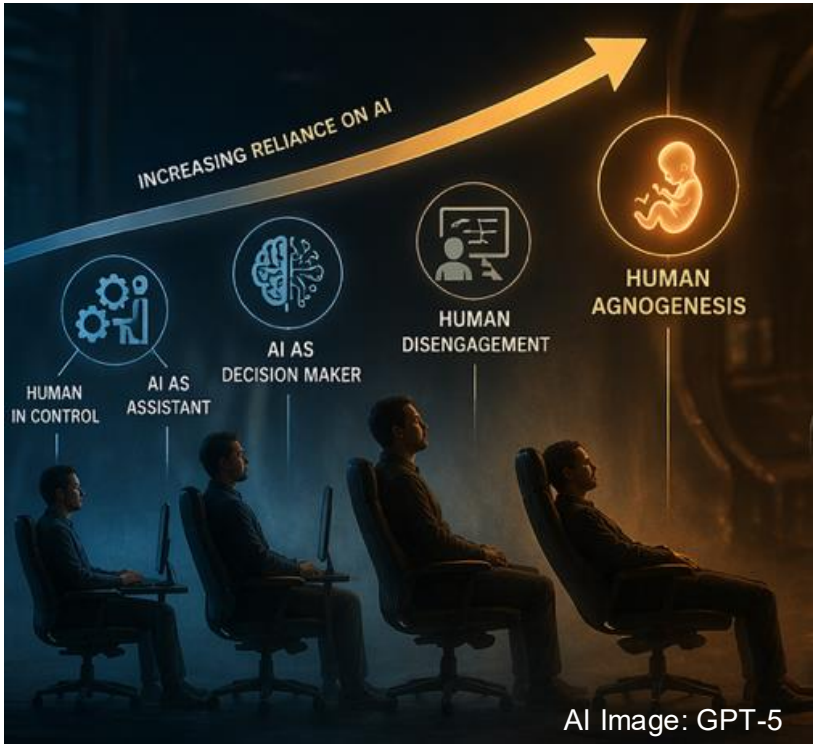
AI Image: GPT-5

*Proctor, R.N., and Schiebinger, L. 2008. Agnotology: The making and unmaking of ignorance. Stanford, CA: Stanford University Press.

OFFICIAL

An Early Reflection…Testing Induced Agnogenesis

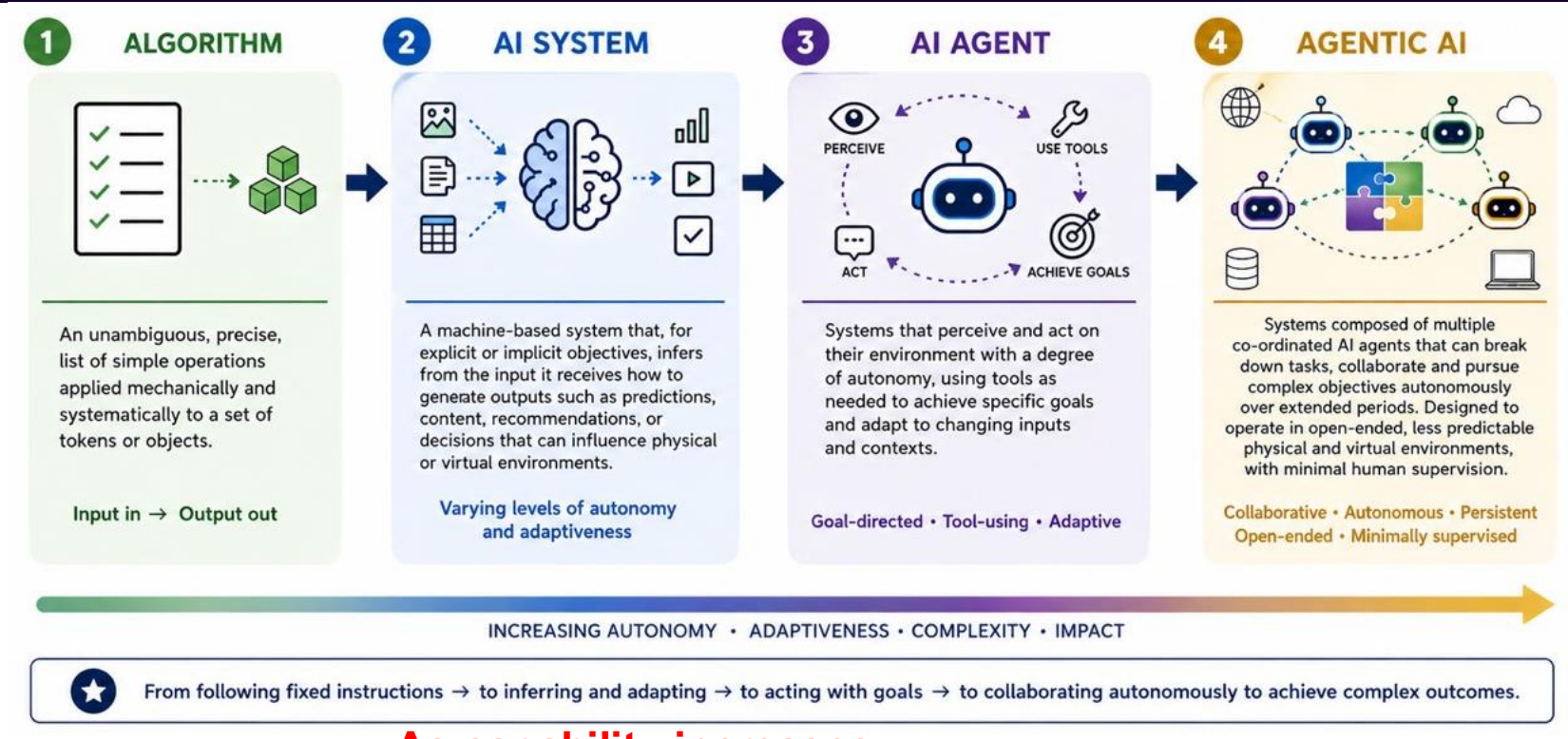
- Some testing benefits such as insight and control are accrued via the act of testing not solely through the tests themselves
- When AI is used in cognitive tasks, without spending time and effort to understand the activity, a gap opens between what is done and what can be defended. This gap widens as further work is built on top, resulting in '*epistemic debt*'*
- Dstl have coined the term '*accidental agnogenesis*' to reflect potential loss of sufficient human insight to maintain effective control over activity



* Unceta, I., Subías-Beltrán, P. & Pujol, O. The epistemic debt of generative AI. *Nat Mach Intell* (2026). https://doi.org/10.1038/s42256-026-01294-w

The greatest risk may be not what we know, or what we don't know, but what we ignorantly believe to have been adequately tested

Getting On The Same Page



As capability increases…

so does the risk of epistemic debt and accidental agnogenesis

AI Image: GPT-5

(based on the Organisation for Economic Co-operation and Development, OECD, definitions)

OFFICIAL

AI In The Context Of Safety Engineering

- AI Safety has a much broader meaning than in traditional safety engineering
- AI can be (amongst things):
  - The product
  - The product developer/manager
  - The product test tool
  - The product test orchestrator/operator/manager
- AI can fill in any role within the software lifecycle EXCEPT one…

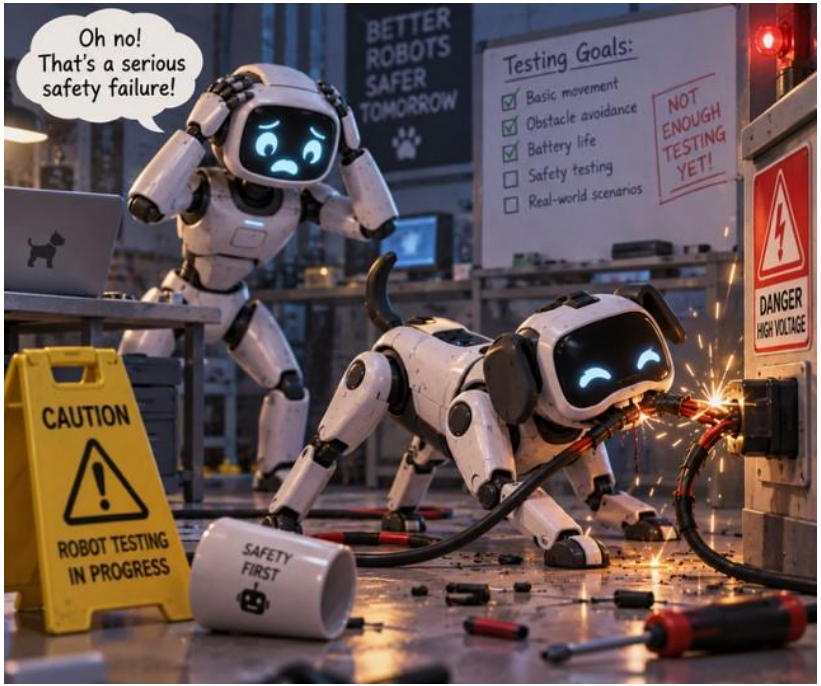
Internet keyword-driven
AI Safety Word Cloud



AI cannot be accountable for anything…
Humans are always accountable, and ignorance is no excuse

AI – What's All The Fuss About?

- AI is probabilistic…
- AI will fail
  - In some systems failure is a success problem???
- Tests will fail
  - But does that make them a failure?
- Safety is a system-level property
  - Failure management may be more important than the failure itself



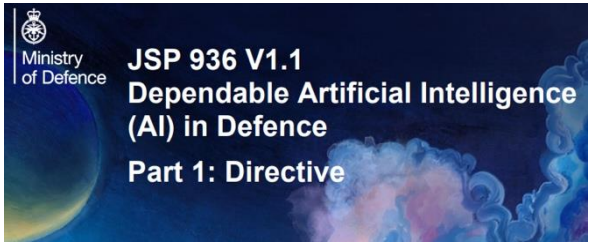
AI Image: GPT-5

The Insightful Test Engineer understands the AI and its context

OFFICIAL

The Systemic Perspective: Policy, Regulation, Standards & Guidance

Dstl co-wrote policy → influenced/co-wrote regulation → wrote guidance



76. Performance targets - especially those relating to the concept of reliability - for AI components **must** be clearly defined for the operating context and demonstrated to a level of confidence that is commensurate with the risk associated with failure. Where AI is complementary to, or replacing, an existing approach, whether that is provided through software, human decision-making or a combination thereof, existing performance targets **must** be considered for continuing acceptability.

77. Where existing performance targets are deemed acceptable then the principle of demonstrating Globally At Least Equivalent (GALE) performance targets may be appropriate.

https://assets.publishing.service.gov.uk/media/6735fc89f6920bfb5abc7b62/JSP936_Part1.pdf

Defence Safety Authority

Military Aviation Authority (MAA)

Regulatory Notice

10 October 2025

MAA/RN/2025/04 – Use of Artificial Intelligence within Systems and Equipment subject to MRP design assurance processes

Issue

1. The continuing maturation and widening adoption of Artificial Intelligence[1] (AI) technologies i digital control systems and decision support tools has led to the requirement to clarify the position with respect to use of such technologies in the Defence Air Environment (DAE).

5. A System or Equipment utilising AI in applications likely to affect overall platform Airworthiness such as propulsion, flight management and flight control systems should demonstrate appropriate consideration of the MOD's five key AI ethical principles[2]. In the context of Air Safety, the principle of reliability is particularly pertinent and as such performance targets for AI components must be clearly defined for the operating context and demonstrated to a level of confidence that is commensurate with the Risk associated with failure. Where AI is complementary to, or replacing, an existing approach, whether that is provided through software, human decision-making or a combination thereof, existing performance targets must be considered for continuing acceptability. Where existing performance targets are deemed acceptable then the principle of demonstrating Globally At Least Equivalent (GALE) performance targets may be appropriate.

https://assets.publishing.service.gov.uk/media/68e9046b1c8b2a3b50690842/MAA_RN_2025_04.pdf

Guidance on the Principle of Globally At Least Equivalent for Performance Targets in AI-based Systems

Report Reference	Date of Issue	Principal Author	Lead Technical Reviewer	Project Manager
DSTL/TR172332 v1	13/10/2025	A Banks	SJ Farmer	E Miller

Introduction

This short paper provides guidance[1] to the reader on the use of the concept of Globally At Least Equivalent (GALE) as highlighted in Joint Service Publication 936 (JSP 936) – Dependable Artificial Intelligence (AI) in Defence [1]. The JSP initially directs that performance targets must be clearly defined for AI as follows:

*"76. Performance targets - especially those relating to the concept of reliability - for AI components **must** be clearly defined for the operating context and demonstrated to a level of confidence that is commensurate with the risk associated with failure. Where AI is complementary to, or replacing, an existing approach, whether that is provided through software, human decision-making or a combination thereof, existing performance targets **must** be considered for continuing acceptability."*

The rationale for this is that without setting performance targets it is difficult for decision-makers to ascertain whether the AI is sufficiently performant for the given application. Quite often, AI is being used to supplement or replace existing tools, techniques or indeed human decision-making and performance targets may already exist. Changes in system use may occur following the introduction of AI technology and as such it is important that extant targets are re-assessed for continuing acceptability.

The JSP then goes on to state:

"77. Where existing performance targets are deemed acceptable then the principle of demonstrating Globally At Least Equivalent (GALE) performance targets may be appropriate."

From a Safety perspective JSP 936 calls out Defence Standard 00-055…

OFFICIAL

Def Stan 00-055* - Embrace The uncertainty

- Based on Hawkins' 4+1**, Def Stan 00-055 sets out the Ministry of Defence's requirements for software safety including AI
- Objectives of Def Stan 00-055 can be summarised as:
  - Define software safety requirements, preserve their intent through decomposition, demonstrate their satisfaction, identify and mitigate hazardous behaviour (including data-related hazards), provide assurance proportional to risk contribution, and address the safety consequences of adaptive behaviour.



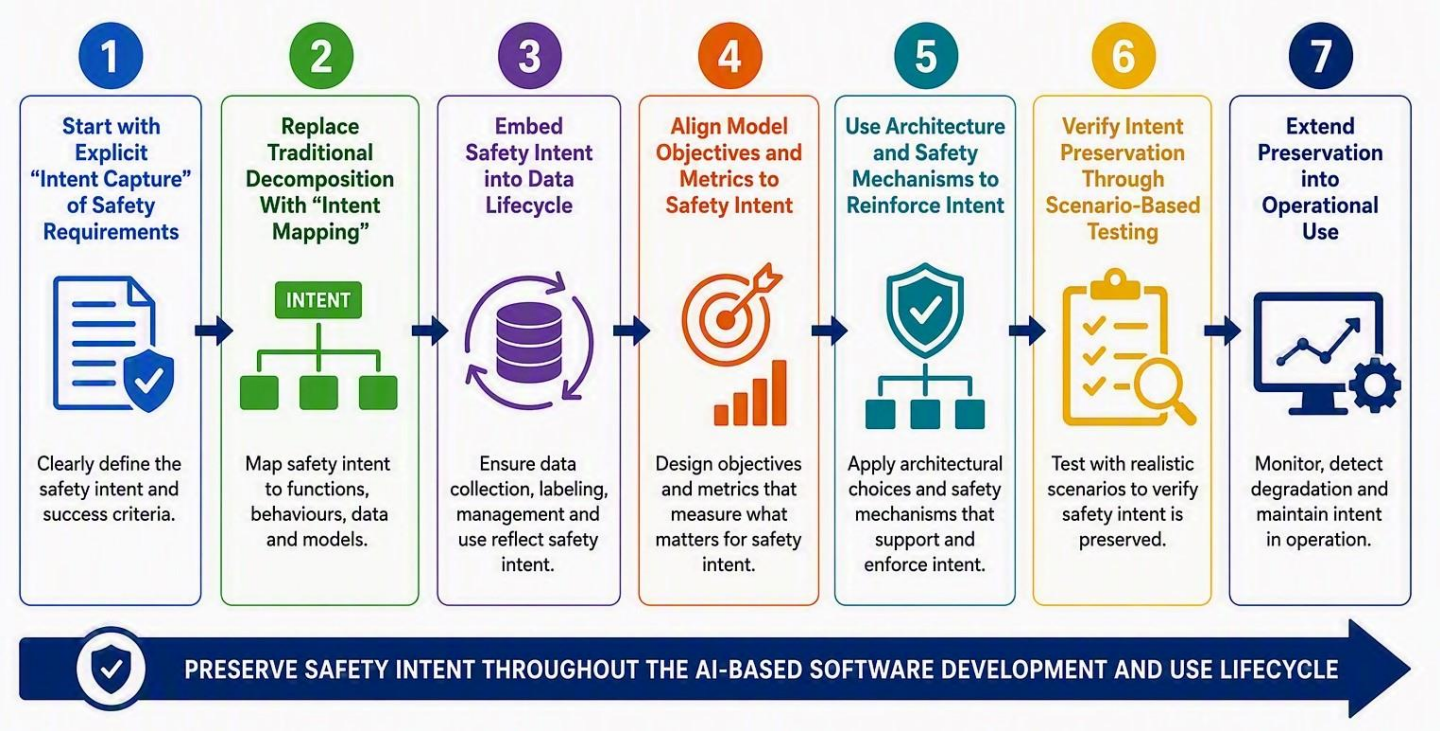
AI in a Def Stan 00-055 context is an assurance challenge – NOT a compliance issue

*Ministry Of Defence. (2021). *Requirements for Safety of Programmable Elements (PE) in Defence Systems. Defence Standard 00-055*

***Hawkins, R., Habli, I., & Kelly, T. (2013). The principles of software safety assurance. 31st International System Safety Conference, . Boston, Massachusetts USA*

Example (Goal 2 of Def Stan 00-055): Understanding The Maintenance Of 'Safety Intent'

- Demonstration of intent preservation is necessarily different for AI



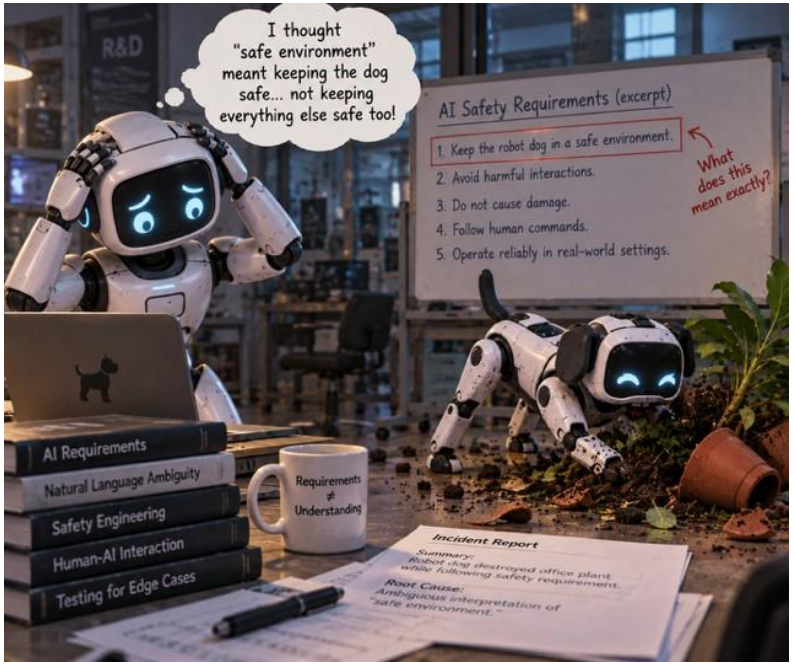
AI Image: Copilot
(based on the currently draft status Dstl AI-Safety Guidance Paper)

Mapping 'intent' is much harder than decomposition tracing… but more insightful

OFFICIAL

Preventing Agnogenesis Via Testing In Our Example (1/2)

- “Intent Capture” of Safety Requirements
  - Review and validation testing of intent statements, hazards, assumptions, and constraints ensures stakeholders share a common understanding of the safety objective, preventing ambiguity from becoming embedded ignorance later in the lifecycle
- Replace Traditional Decomposition with “Intent Mapping”
  - Traceability and consistency testing between high-level safety intent and decomposed functions verifies that safety meaning has not been diluted, altered, or lost during allocation to subsystems
- Embed Safety Intent into the Data Lifecycle
  - Dataset audits, coverage analysis, bias assessment, and data lineage verification confirm that the data reflects the original safety concerns, preventing ignorance arising from unrepresentative, incomplete, or poorly labelled datasets



AI Image: GPT-5

OFFICIAL

Preventing Agnogenesis Via Testing In Our Example (2/2)

- Align Model Objectives and Metrics to Safety Intent
  - Objective-function validation and metric suitability testing demonstrate that the model is being optimised for the intended safety outcomes rather than surrogate measures that could obscure safety-relevant behaviour
- Use Architecture and Safety Mechanisms to Reinforce Intent
  - Architectural assurance testing, safety mechanism verification, fault-injection, and fail-safe testing confirm that system design actively preserves safety intent even when components fail or behave unexpectedly
- Verify Intent Preservation Through Scenario-Based Testing
  - Scenario-based, edge-case, and operational-design-domain testing evaluates whether the system's behaviour remains consistent with the original safety intent across realistic and challenging conditions, preventing false confidence based on limited test evidence
- Extend Preservation into Operational Use
  - Runtime monitoring, drift detection, incident analysis, and performance assurance testing identify emerging deviations from safety intent in deployment, preventing the gradual accumulation of ignorance as the operating environment evolves



AI Image: GPT-5

OFFICIAL

Key Takeaway

[dstl] Delivering Mission Success

"The challenge before us is to evolve from proving systems work to discovering what we may have missed. In an AI-enabled world, the test engineer's highest duty is not evidence generation, but the prevention of ignorance through the relentless pursuit of evidence-based insight."

OFFICIAL

[dstl]
Delivering
Mission Success

Discover more

OGL

The contents include material subject to © Crown copyright (2026), Dstl.

This information is licensed under the Open Government Licence v3.0. To view this licence, visit https://www.nationalarchives.gov.uk/doc/open-government-licence/

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.
Any enquiries regarding this publication should be sent to:
centralenquiries@dstl.gov.uk [other author(s) / authoring organisation(s)].