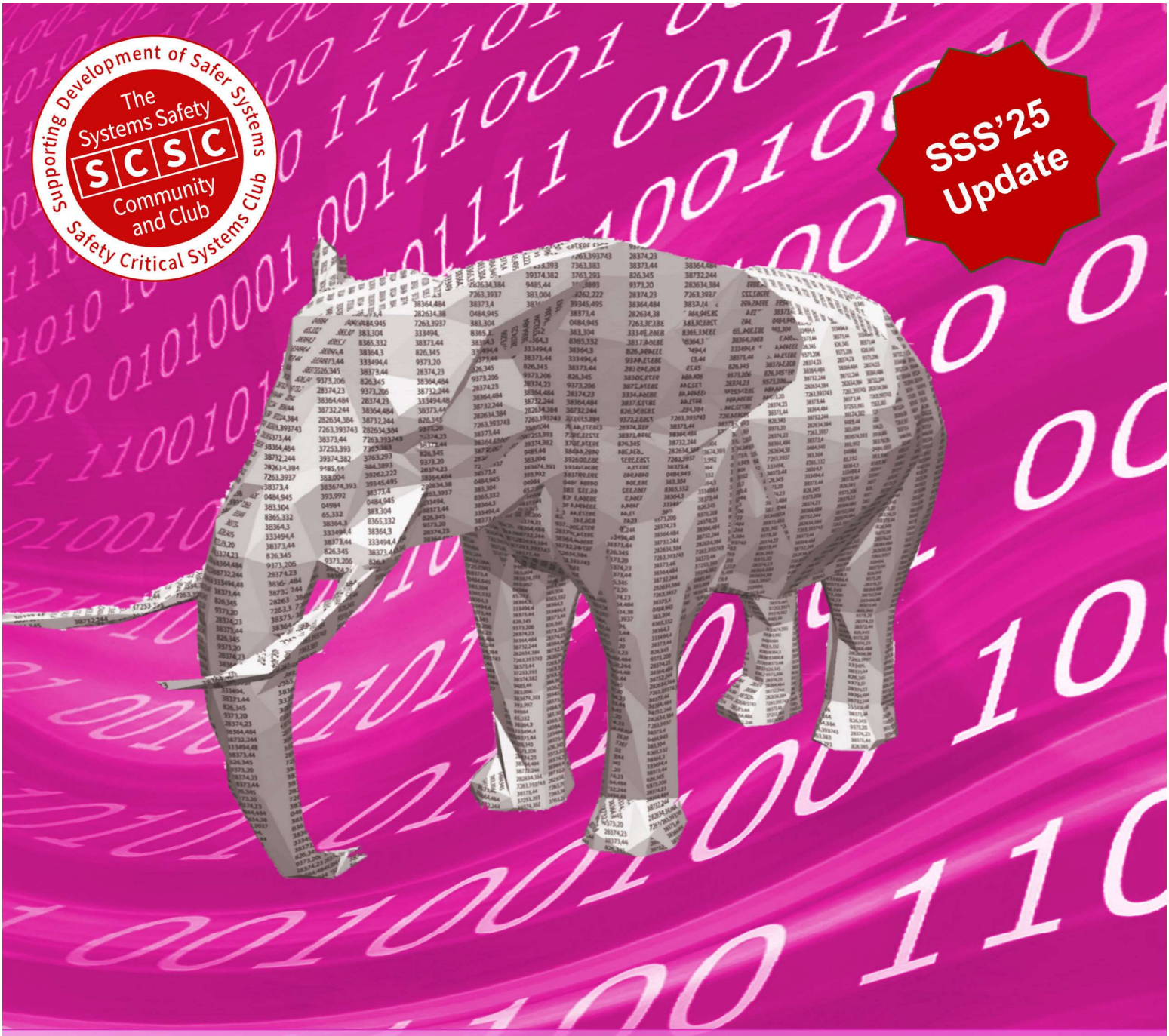


Data Safety Guidance

Version 3.7

The Data Safety Initiative Working Group (DSIWG)

SCSC-127J

This page is intentionally blank

SCSC Publication Number: SCSC-127J

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA. You are free to share the material in any form and adapt the material for any purpose providing you attribute the material to the Safety-Critical Systems Club (SCSC) Data Safety Initiative Working Group, reference the source material, include the licence details above, and indicate if any changes were made. See the license for full details.

This document was prepared using the LaTeX 2_ε typesetting system.

Editing and typesetting by Mark Templeton, supported by Tim Rowe.

Cover design by Paul Hampton.

The Safety-Critical Systems Club (SCSC) is the professional network for sharing knowledge regarding safety-critical systems. It brings together:

- engineers and specialists from a range of disciplines working on safety-critical systems in a wide variety of industries;
- academics researching the arena of safety-critical systems;
- providers of the tools and services that are needed to develop the systems; and
- the regulators who oversee safety.

Through publications, seminars, workshops, tutorials, a web site and, most importantly, at the annual Safety-critical Systems Symposium (SSS), it provides opportunities for these people to network and benefit from each other's experience in working hard at the accidents that don't happen. It focuses on current and emerging practices in safety engineering, software engineering, and product and process safety standards.

This document was written by the Data Safety Initiative Working Group (DSIWG), which is convened under the auspices of the SCSC. The document supports the DSIWG's vision, which is to have clear guidance that reflects emerging best practice on how data (as distinct from software and hardware) should be managed in a safety-related context. This update takes account of the consensus that a process-based guidance document will complement existing safety management processes, making it more usable. It was formally released at SSS'25, 4–6 February 2025 details of which may be found at https://scsc.uk/e1099

Comments on this document are actively encouraged. These can be emailed to:

comments@data-safety.scsc.uk.

Alternatively, a comments submission form is available at:

data-safety.scsc.uk/comments.

While the authors and the publishers have used reasonable endeavours to ensure that the information and guidance given in this work is correct, all parties must rely on their own skill and judgement when making use of this work and obtain professional or specialist advice before taking, or refraining from, any action on the basis of the content of this work. Neither the authors nor the publishers make any representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability or availability with respect to such information and guidance for any purpose, and they will not be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever (including as a result of negligence) arising out of, or in connection with, the use of this work. The views and opinions expressed in this publication are those of the authors and do not necessarily reflect those of their employers, the SCSC or other organizations.

Data Safety Guidance

The Data Safety Initiative Working Group [DSIWG]

February 2025

This page is intentionally blank

Change History

Version	By	Status	Date
1.0	The DSIWG Team	First draft for external review	31-JAN-2014
1.1	The DSIWG Team	(Internal edition for DSIWG use only)	09-DEC-2014
1.2	The DSIWG Team	For publication at SSS'15	23-JAN-2015
1.3	The DSIWG Team	For publication at SSS'16	29-JAN-2016
2.0	The DSIWG Team	For publication at SSS'17	30-JAN-2017
3.0	The DSIWG Team	For publication at SSS'18	26-JAN-2018
3.1	The DSIWG Team	For publication at SSS'19	01-FEB-2019
3.2	The DSIWG Team	For publication at SSS'20	11-FEB-2020
3.3	The DSIWG Team	For publication at SSS'21	09-FEB-2021
3.4	The DSIWG Team	For publication at SSS'22	08-FEB-2022
3.5	The DSIWG Team	For publication at SSS'23	07-FEB-2023
3.6	The DSIWG Team	For publication at SSS'24	13-FEB-2024
3.7	The DSIWG Team	For publication at SSS'25	04-FEB-2025

Changes Since the Last Edition

The main changes in this edition are:

- the inclusion of a process flow diagram at Section **??**;
- the inclusion of an appendix (Appendix 5) on AI and autonomy; and
- the inclusion of an appendix (Appendix 8) on the RADISH tool.

The inclusion of the new Appendix 5 and Appendix 8 means that subsequent appendix numbers have changed.

Hyperlinks (in the electronic edition) to abbreviations, acronyms and defined terms have been substantially expanded and further abbreviations, acronyms and defined terms have been added.

The body of text the guidance, but not yet the appendices, has been edited for clarity and consistency.

Discursive definitions of terms and abbreviations not used in the document have been removed, and new discursive definitions and definitions have been addded.

The definition of hazard has been expanded to include harm, not just accidents.

An inconsistency in the worked example regarding whether there is a contract for the system has been corrected.

To assist users of earlier 3.x versions of the guidance in ensuring that their existing data safety arguments have not been impacted by this update, a version of this document is available which has been annotated with change bars. To avoid clutter, minor changes that should not affect the meaning of the text have not been marked with change bars. The annotated version is available at http://scsc.uk/scsc-127J

Future work

MCA Ltd has continued to work with the DSIWG to develop a prototype software tool to assist in the automation of the processes described in this guidance document. A working version of the tool has been developed and organizations that could benefit from the use and further development of the tool are urged to contact MCA at Mission Critical Applications Limited (radish@mca-ltd.com).

A number of improvements to the guidance are currently planned. These improvements are intended to clarify the application of the data safety process and include:

- further detail on the assurance of communications and data flows,
- data safety considerations associated with distributed datasets and Blockchain,
- addition of new treatments to the tables in **??**,
- review of the tables of treatments, with the aim of making them easier to use,
- further explanation of some treatments, where their use or benefit is not immediately apparent,
- reordering of parts of the document to improve readability, especially as regards likelihood,
- further detail on tool assurance,
- harmonisation of language and guidance on how organizations may expand the tables to incorporate their own internal processes.
- guidance on the application of the data safety culture questionnaire,

Several of these changes are likely to cause parts of the document to be re-ordered – they have therefore been deferred to the next major update, in version 4.0 of the guidance.

If you or your organization are interested in learning more about the work of the DSIWG or joining any of the sub-groups, please visit the SCSC website, where more information including contact details may be found on the "Working groups" section of the site.

Related working groups

The SCSC sponsors initiatives to develop methods and techniques through a number of working groups. These groups each address safety aspects peculiar to their domain, including data aspects when appropriate. The current list of working groups includes:

- Assurance Cases,
- Security Informed Safety,
- Safe AI Working Group,
- Safer Complex Systems.

This page is intentionally blank

Foreword

> *Data is here. Data is growing. Data is causing harm.*

Data is here: Data is becoming ever more important in our lives: influencing, managing and even controlling many critical aspects. The use of artificial intelligence (AI) systems is a new, exciting, but potentially hazardous use of data. Large language model (LLM) based systems are trained on vast amounts of data, and it is this data which enables them to be useful.

Some of this data is related to our personal safety and well-being. Consider, for example, the importance of data defining the layout of railway signals, data that indicates the position of underwater obstructions in nautical channels or data that is used to train a vision recognition system to detect tumours in medical images. Organizations now make significant decisions (including safety-related decisions) based solely on data held in systems. Hence, organizations need to safely manage, control and process their data. In particular, they must actively manage key data properties that preserve safety.

Data is growing: There are at least two reasons why the use of data has grown and, equally important, why it is expected to continue to grow. The first relates to the rapid expansion of the area loosely termed "Big Data", including the use of large datasets to support machine learning and AI applications. The second is the growing use of systems of systems, where data is the lifeblood that connects together disparate elements and allows a cohesive capability to be built. Put simply, the need to address data-related issues is a pressing problem and will continue to be so.

Data is causing harm: Strictly speaking, data can neither cause nor prevent harm. However, mistakes in data, or the inappropriate use of data, within safety-related systems have been factors in a number of documented accidents and incidents. Examples include aircraft attempting to take off from the wrong runway (and consequently crashing), ships running aground, and patients being exposed to higher than planned doses of radiation.

Against this background, the DSIWG was established under the auspices of the SCSC. The DSIWG's aim is to develop clear, cross-sector guidance that reflects emerging best practice on how data (as opposed to software or hardware) should be managed in a safety-related context. For the most part, this guidance is based on well-established techniques, and it has been designed to be compatible with current safety standards and to integrate with existing safety management systems. What is new, however, is the explicit and relentless focus on data, making it a "first-class citizen" within system safety analyses. Because of this focus, this guidance should help organizations identify, analyse, evaluate and treat data-related risks, thus reducing the likelihood of data-related issues causing harm in the future.

This page is intentionally blank

Quick Start Guide

> *Data really powers everything that we do.*
> ***Jeff Weiner***

This section provides a single-page introduction to data safety guidance (DSG). For first-time readers this should help place individual sections within an appropriate context. It should also help returning readers quickly navigate the document's contents.

- Systems are changing. The role of data is becoming more prominent. Hence, data needs to be considered as a "first-class citizen" in system safety analyses. This will help mitigate organizational and system-level risks associated with the use of data.
- A data safety management process has been developed. This is based on four phases:
  - establish context;
  - identify risks;
  - analyse risks; and
  - evaluate and treat risks.
- The underlying principles and an overview of the process are described in **??**.
- Normative definitions and abbreviations are described in chapter 1.
- The objectives associated with, and the outputs produced by, each phase are described in **??**.
- The activities of each phase (and associated tailoring information) are described in **??**.
- Additional guidance information for each phase is described in **??**.
- A worked example is provided in chapter 2.
- A collection of appendices provide more detail, including:
  - A discussion illustrating how the underlying principles link to the objectives (Appendix **??**);
  - An organizational data risk (ODR) assessment questionnaire (Appendix **??**);
  - A data safety culture questionnaire (Appendix **??**);
  - A questionnaire to help assess the data maturity of a supplier (Appendix **??**);
  - A list of data categories (Appendix **??**);
  - A collection of hazard and operability study (HAZOP) guidewords (Appendix **??**);
  - The suggested contents of a data safety management plan (DSMP) (Appendix **??**);
  - A summary of accidents and incidents in which data was potentially a causal factor (Appendix **??**);
  - A discussion of topics loosely related to system lifecycles (Appendix 3);
  - Considerations regarding machine learning (ML) (Appendix 4);
  - A discussion of the risks of AI and autonomy (Appendix 5);
  - An introduction to the concepts of both dark and dazzle data (Appendix 6);
  - The concepts of black swan, dragon king, perfect storm and Pudding Lane data (Appendix 7).
  - Considerations for the assurance and qualification of data-handling tools (Appendix **??**).

- An introduction to the RADISH tool, that has been developed to assist in the application of the guidance within this document (Appendix 8).
- Issues that may arise when migrating, porting, importing or exporting data (Appendix **??**).
- Some of the data issues that made management of the Covid-19 virus difficult (Appendix 9);
- Examples of ways that data safety assurance levels (DSALs) may be customised, with particular focus on likelihood (Appendix **??**);
- Lists of acronyms, definitions and glossary entries (Appendix 9.5); and
- A collection of references (Appendix 9.7).

This page is intentionally blank

This page is intentionally blank

Contents

List of Tables

This page is intentionally blank

List of Figures

This page is intentionally blank

0.1 Introduction (Informative)

> *We're entering a new world in which data may be more important than software.*
> ***Tim O'Reilly***

This page is intentionally blank

0.2 Aim and Scope

This guidance document aims to:

- describe the data safety problem;
- provide methods for identifying and analysing levels of risk; and
- recommend methods and approaches for evaluating and treating those risks.

It has been written for a wide readership. Its target audience is all those who have an interest in or a responsibility for safety-related data within systems, including managers, developers, safety engineers, assurers (including independent safety auditors), regulators, and operators.

The document is also intended to cover a number of different sectors. It identifies a wide spectrum of safety-related data that exists in many forms within systems, from specification and requirements data to maintenance and disposal data, and everything in between. In particular, this document is not just concerned with numerical or well-structured data used during system operation.

While they are considered mature enough to be useful, the contents of the document represent current thoughts on what is a complex and evolving area. Furthermore, to allow it to be produced within a reasonable timescale, this edition focuses on key items. It is not intended to be exhaustive. For example, this guidance document does not consider issues relating to staff competence or organizational structure.

This page is intentionally blank

0.3 Intended Relationship to Other Documents

This document is intended to be used as a supplement to existing standards and norms that are relevant to the scope of the work being undertaken. It may be used to provide a deeper insight into the risks that data poses to the project team's outputs, allowing them to produce credible improvements to the safety argument. Where a standard or norm sets out specific data-related objectives then, unless agreed otherwise with the regulator or safety duty holder, they shall take precedence over the guidance provided herein.

In the longer term, the hope is that future standards and norms will take up relevant concepts, approaches and methods from those in this document. The DSIWG also hopes that organizations will include the concepts, approaches and methods in their own safety management processes.

This page is intentionally blank

0.4 Normative, Informative and Discursive Text

Three types of text are used within this guidance document:

Normative text, which is prescriptive. Typically, this text is restricted to describing objectives and outputs.

Informative text, which is descriptive text that is closely linked to the normative text. Typically, this text provides a suggested way by which compliance with the normative text may be achieved, but alternative means of compliance are possible.

Discursive text, which contains discussions that are relevant to the general topic of data safety, but which are not closely linked to the normative text. A discussion on the relationship between data and software is an example of such text. Descriptions of historical incidents and accidents are another.

Each section and appendix of this guidance document contains a single text type. The relevant type is indicated in the section or appendix title.

This page is intentionally blank

0.5 Compliance

There may be occasions when it is desirable or necessary to make a claim of compliance against the objectives listed in this document. Such a claim may be required, for example, if this document is explicitly included as a normative reference from a formal standard. Alternatively, it may be required as part of an organization's internal processes.

To facilitate compliance claims, the following terminology is used within the normative parts of this guidance document:

SHALL denotes items where evidence of compliance must be provided in order to claim compliance with this guidance document.

SHOULD denotes items where, in some circumstances, there may be valid reasons for not complying with a particular item. The full implications of non-compliance must be understood, documented and approved in order to claim compliance with this guidance document.

MAY denotes items that are optional. These may be advantageous in some circumstances but not in others. Organizations are free to adopt any approach to these items without the need for further justification.

The terms have their normal English meanings in discursive and descriptive sections.

This page is intentionally blank

Chapter 1

Definitions (Normative)

> *When I use a word,' Humpty Dumpty said in rather a scornful tone, 'it means just what I choose it to mean – neither more nor less.*
> ***Lewis Carroll***

This document is incomplete. The external file associated with the glossary 'normative' (which should be called `Vol3.normative-gls`) hasn't been created.

Check the contents of the file `Vol3.normative-glo`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:
  ```
  \usepackage[automake]{glossaries-extra}
  ```
- Run the external (Lua) application:
  ```
  makeglossaries-lite.lua "Vol3"
  ```
- Run the external (Perl) application:
  ```
  makeglossaries "Vol3"
  ```

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed. A more comprehensive list of definitions, including descriptive definitions, is included at chapter 9.5.

This page is intentionally blank

Chapter 2

Worked Example (Informative)

> *There is a forest of data and we need to create a path through.*
> ***Tom Adams***

This page is intentionally blank

2.1 Purpose

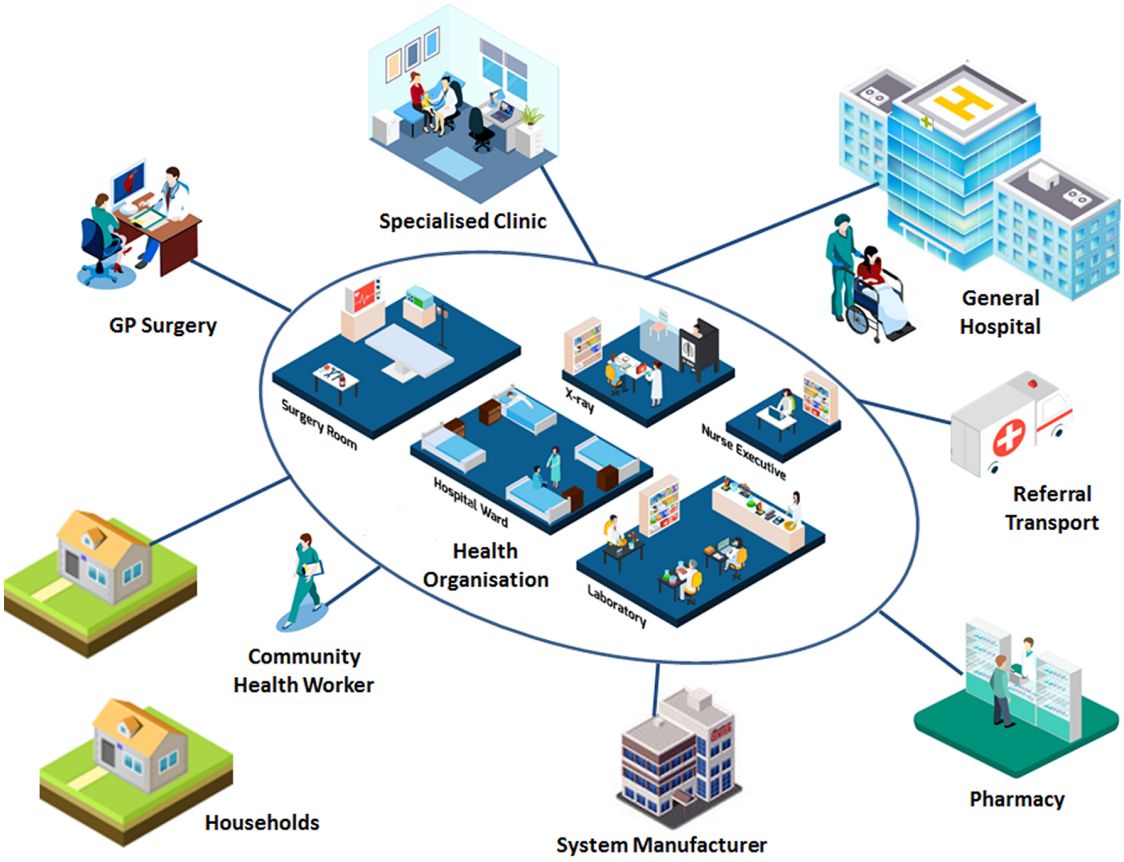
This section provides a worked example of applying the DSG to a hypothetical system in the healthcare sector. Although some aspects of the example have been simplified, it is intended to be sufficiently realistic to allow key features of the guidance to be illustrated.

This page is intentionally blank

2.2 Establish Context

2.2.1 Background

A manufacturer is building a new integrated health and social care system to support holistic care for community health services. The system will support clinical workflows for aspects such as referrals, tracking clinical encounters, appointment scheduling, outcome measures through to letter and report generation. The development will follow a typical development lifecycle as a series of phases: business modelling, requirements, analysis and design, implementation, test, and deployment.



The system is being targeted to meet the requirements of a health organization that is procuring a solution to help their clinicians maintain a high level of quality of care in the face of increasing volumes of patients and pressure to reduce staff costs.

Fundamental to fulfilling these requirements is establishing the context within which data is used in system development, enhancement, introduction, integration or operation. This should establish the risk appetite: essentially, how much effort is to be devoted to making data risks as low as practicable. In turn, this will inform the nature and scope of assessments that are to be conducted during system development and its introduction into operational service. To meet the 'establish context' objectives set out in the guidance the following activities are recommended:

- describe the organizational context;
- describe the system context;
- plan the assessment; and
- identify data artefacts.

The manufacturer decides to use an ODR assessment form to understand in broad terms the level of risk it will have to manage in developing and supporting this system. This will allow the manufacturer to describe the organizational context and describe the system context.

System Manufacturer

2.2.2 ODR Assessment

The manufacturer considered each of the questions within the ODR form from**??** of this document. The questions in the ODR and the manufacturer's assessments were:

Q1:	How severe could an accident be that is related to the data? Could it be caused directly by the data?

Failings in the system could give rise to non-optimal treatment plans for a patient that might delay detection of a more serious condition or prolong the recovery for a known condition. The system could not directly cause an accident, however, and there are other people and systems in play involved in checking data. On balance, this question is assessed as **1c**; Score **4**.

Q2:	What would be the impact on the organization, client or public if an accident occurred related to the data?

Unfortunately, accidents in the health domain are relatively frequent. There are many injuries and deaths attributed to medical errors, but these are largely tolerated by the public and grievances are usually settled by financial settlements through the courts. The manufacturer believes their contractual arrangements mean that the health organization would be liable for any claims even if they were attributed to an error in the manufacturer's system's handling of data. On balance, this question is assessed as **2b**; Score **2**.

Q3:	How much responsibility does this organization have for data safety?

The manufacturer is responsible for building the system in compliance with DCB0129 [**?**] and so is responsible for executing the associated safety management system to manage risk. The manufacturer, however, plans to sell the product with a condition of use that places end responsibility for patient safety on the client. On balance, this question is assessed as **3b**; Score **2**.

Q4:	What legal and regulatory environment will this work be subject to?

The work will be contracted under UK law and subject to the data coordination board (DCB) standards for health IT Systems. However, there is no regulator currently empowered to intervene in the delivery of healthcare systems, so the standards are not currently enforced through law. Instead, the health organization will make compliance with the standards a contractual requirement. On balance, this question is assessed as **4c**; Score **4**.

Q5:	How mature is this organization regarding data safety?

The manufacturer has a good understanding of data as a source of safety risk. Many of their systems are data intensive to support clinical decision making. There is good support and funding for the identification and treatment of data-related risks. On balance, this question is assessed as **5b**; Score **2**.

Q6:	How widely used is the data and who by?

The data will be used in multiple clinical settings and by many clinicians and other support staff. There are several data supply chains and public web access to data. On balance, this question is assessed as **6c**; Score **4**.

Q7:	What is the scale, sophistication and complexity of the data and its manipulation?

The data is complex and although transmitted through industry standard data structures these require knowledge of the associated abstract clinical data model. Some data manipulation is required to map between different encodings of data held in the various heterogeneous systems. Some legacy systems transfer data in unstructured format. On balance, this question is assessed as **7c**; Score **4**.

Q8:	How well defined and understood are the boundaries and interfaces for this data scenario?

The boundaries of the supply are well understood and although the interfaces are complex and mixed formats, these will be defined and agreed formally through interface control documents (ICDs). Most of the integrating systems are established commercial off-the-shelf (COTS) based systems, but some of the legacy systems still need to be investigated and working assumptions have been made by the manufacturer. On balance, this question is assessed as **8c**; Score **4**.

The final score is **26**, which corresponds to **ODR2**. The manufacturer therefore concludes that there is low to medium risk that loss of properties of data in the system can contribute to or give rise to harm. The manufacturer has an internal policy for engagements based on the ODR level that dictates how the organization will plan the assessment. This policy dictates the amount of proportional effort the

manufacturer needs to spend on safety data management and the level of rigour to be employed. In this case, the policy dictates, amongst other requirements, that a separate section covering safety data management is required in its clinical risk management plan.

Next, the manufacturer identifies data artefacts that are potential sources of safety hazards. The manufacturer knows that the safety dependency of data is dictated by the context in which it is used so it also develops an understanding of when in its process lifecycle the data will be used and relied upon. The manufacturer plans to build an early prototype to show to clients to help elicit requirements definition. To support this, the manufacturer plans to create a test dataset that comprises a typical range of scenarios that the system will encounter. This form of data is identified as verification data. The system also needs to be configured to support deploying health organizations' policies. This data is infrastructure data and, for the prototyping phase, the manufacturer plans to use largely default values.

In later phases, when the system functionality is specified and the system is being built, the manufacturer plans to create a test dataset that will be key to demonstrating the correct functioning of the system and hence acceptance by the deploying health organization. This still involves the use of verification data and infrastructure data but there will be far greater dependency on these datasets than the prototyping case. The manufacturer therefore documents in the clinical risk management plan the planned use of each of the data categories during the entire delivery lifecycle.

The procuring health organization will have a different perspective of the IT system that it will deploy into its organization. It will already have many integrated systems in live operation and as part of establishing the context for the system's deployment they will need to consider many different types of datasets:

Infrastructure data: how the system will be configured in the specific environment;

Verification data: the test datasets to be used to support certain deployments such as integration testing and training; and

Dynamic data: the data entered or fed into the system and the data presented to the user, generated in the form of reports or data passed to other systems.

The health organization decides to complete an ODR assessment so it can describe the organizational context and describe the system context. The scoring is similar to the manufacturer with the following notable differences:

Q2:	What would be the impact on the organization, client or public if an accident occurred related to the data?

The health organization would bear the brunt of any publicity and litigation in the event of an accident and so assesses this question as **2c**, Score **4**.

Q3:	How much responsibility does this organization have for data safety?

The health organization is responsible for deploying systems in compliance with DCB0160 [**?**] and is ultimately responsible for patient safety. The health organization assesses this as **3e**, Score **12**.

Q5:	How mature is this organization regarding data safety?

The health organization has only recently acquired the expertise to apply DCB0160 and is still developing its capability. Safety data management is new to the organization and it anticipates some resistance from senior management to the expenditure incurred in rolling out a DCB0160 compliant safety management system. This question is assessed as **5d**, Score **7**.

The resulting score for the health organization is **43**. This is **ODR3**: medium to high risk. The health organization aims to plan the assessment through a clinical risk management plan. This plan defines the organization and system context in more detail and lays out the planned activities for identifying, evaluating and treating data safety related risks.

As with the manufacturer, the health organization needs to identify data artefacts that are potential sources of safety hazards and understand the context of their use in the procurement / deployment lifecycles of the health organization. Post acceptance, the procuring health organization plans to run a series of user training sessions for clinicians. Once users are trained, the system will be integrated into live operations. The health organization identifies the infrastructure, verification and dynamic data categories to be used during these phases. The health organization also realises that the system will form part of a data supply chain, because a number of external organizations and departments within their own organization engage in the procurement and use of safety-related data. For example, it will receive referral data from a number of other general practitioner (GP) systems, it will receive outcome measures from hospitals, and it will receive clinical data acquired from remote workers visiting patients in the community and from the patients themselves using the system's online portal. The system also produces data for other external systems such as electronic prescriptions for pharmacies.

The health organization sees that by using the new system it will become a commissioning user, as it will require and be a consumer of data from a variety of sources: GP systems, hospital systems, systems used by remote workers in the community, and the system's portal capturing data entered by the patients themselves. All of these act as data provisioners. Those health care professionals patients gathering patient data through physical inspections and measurement are the data acquirers.

The health organization defines in its clinical risk management plan the data supply chain relevant to the system including the roles and interfaces involved. This will therefore show where there are dependencies on dynamic data used and produced by the system.

Questions the health organization will need to address when establishing the context are:

- have all the dependent interfaces been identified?
- have the roles of commissioning user / data provider / data acquirer been established and acknowledged?
- what service level agreements or contracts exist for the delivery of the data?
- what level of assurance do data providers / data acquirers provide for their data?

This page is intentionally blank

2.3 Risk Identification

The manufacturer aims to carry out the following activities to meet the guidance objectives for the risk Identification phase:

- review the general, historical perspective;
- conduct a top-down approach;
- conduct a bottom-up approach; and
- update planning documents.

Before embarking on any hazard analysis, the manufacturer ensures that stakeholders review the general, historical perspective. This takes the form of a refresh briefing to raise awareness of issues that are specific to data such as ageing, biasing and defaults.

The manufacturer of the health IT System decides that during the prototyping phase there is little safety dependency of the test and configuration data sets as no clinical decisions will be made based on their content; the data is simply being used to support the elaboration of requirements.

In later phases however, when the system functionality is specified and the system is being built, the manufacturer will want to create a test dataset that will be instrumental in demonstrating the correct functioning of the system. This still involves the use of verification data and infrastructure data, but there is far greater dependency on these datasets than the previous case. For example, if the verification or configuration data is not sufficiently diverse or if it insufficiently models real-world scenarios, it is possible that erroneous and unsafe functional behaviour is present in the system during live operation despite the system having passed factory and site acceptance testing.

To analyse the risks in more detail, the manufacturer uses a top-down approach and a bottom-up approach. In the first approach it considers each of the system functions (such as clinical screens) and analyses where there is a dependency on data and what properties need to be preserved. In the second, as much of the functionality is driven by data flows in and out of the system, the manufacturer also looks at specific data flows and assesses the impact of loss of properties for the data in those flows.

On completion of the risk identification phase the manufacturer updates planning documents such as the clinical risk management plan, to reflect the outcome of the analysis.

The health organization will likewise need to conduct risk identification relevant to their deployment context. Hazards arising from data sources that are to be delivered into the new system from existing systems need to be assessed for data risks. As with the manufacturer, a briefing to stakeholders to review the general, historical perspective is first conducted to cover generic data safety issues but also to highlight lessons learned from previous accidents and incidents that have occurred in the health organization itself.

As with the manufacturer, both a top-down approach and a bottom-up approach are adopted. From the health organization's perspective, one key focus for hazard identification is in the use of dynamic data, i.e., data that will be delivered into the new system from existing system data sources and the data presented to the user. For the interactions identified in the supply chain, the health organization needs to consider the risks associated with loss of properties of the data it will receive. Questions the health organization will need to consider and address more formally in the clinical risk management plan are:

- which datasets or data items being received from other systems have data properties (such as timeliness, completeness, consistency, fidelity / representation etc.) that are significant to patient safety?
- what data presented to the user has data properties (such as availability, format, resolution, etc.) that are significant to patient safety?
- what existing barriers or mitigations (physical, technical, procedural) exist to reduce the risk of loss of data properties? and
- will any existing barriers be lost as a consequence of the new system?

On completion of the Risk Identification phase the health organization will update planning documents such as the clinical risk management plan to reflect the outcome of the analysis.

2.4 Risk Analysis

In this phase identified hazards are assessed to determine their likelihood and severity. To meet the guidance objectives the following activities are carried out:

System Manufacturer

- establish DSALs; and
- analyse DSALs as part of system safety activities.

The manufacturer will establish DSALs by considering cases where the use of specific categories of data could give rise to hazards. In the prototyping phase, the manufacturer sees no use of the data that can give rise to credible clinical risk and assesses the DSAL for that dataset as **DSAL0**.

In the second phase of the development lifecycle, where verification data and infrastructure data is being used to demonstrate the correct functioning of the system, the manufacturer considers that loss of any of the data properties of integrity, completeness, consistency, continuity, format, accuracy, resolution, timeliness, availability, fidelity / representation, sequencing, intended destination / usage, Goldilocks, and explainability of this data could give rise to hazards.

For example, if the verification dataset selected is not representative of the eventual diversity experienced in practice (loss of fidelity / representation), then it is possible that the system may contain latent software errors that could give rise to harm. However, the manufacturer acknowledges that the system will be subject to further testing and trials in the clinical setting and so there will be other opportunities to detect errors in the system. Overall:

- the likelihood of the data use gives rise to an accident is **Medium** as other systems and processes are in place that would detect errors; and
- the severity is **Moderate**; failings in the system could give rise to non-optimal treatment plans for a patient that might delay detection of a more serious condition or prolong the recovery for a known condition.

The manufacturer therefore assesses these data categories as **DSAL1** in this particular context of use.

The manufacturer takes care to analyse DSALs as part of system safety activities by documenting these assessments along with other hardware and software safety considerations, for example those arising from DCB0129, in the clinical risk management plan.

From the health organization's perspective, the main focus for risk assessment to establish DSALs is in the use of dynamic data. For the interactions identified in the supply chain the health organization needs to consider the risks associated with loss of properties of the data it will receive and present to the user. Questions the health organization will need to consider and address more formally in the clinical risk management plan are as follows:

Health Organisation

- how likely is it that there would be a loss of the given data property?
- how would such a loss of a data properties be detected?
- how would such as loss be isolated to prevent further risks of harm? and
- hhat recovery action would be required to resolve the issue to maintain patient safety?

Concerns were raised about the vulnerability of the system to inadvertent data overload from the large number of potential data sources, and possibly from malicious activity. This led to the addition of the Goldilocks property to the list of data properties applicable to the live system.

In considering the receipt of outcome measures data received from a clinic or hospital, the health organization considers that it is likely that some credible errors would not be readily detected by their new system; if the hospital system confused a result or there were errors in the precision of data then there would be few chances to catch these once received by the system.

- The health organizations assesses the likelihood of this loss of property as **High**; and
- The impact of such errors, although not realistically likely to lead to death, could result in delays to treatment that could result in serious injury and hence **Moderate** impact.

The data received from this data source is therefore classed as **DSAL2** in this particular context of use.

The health organization takes care to analyse DSALs as part of system safety activities by documenting these assessments along with other hardware and software safety considerations (e.g., arising from DCB0160 requirements) in the clinical risk management plan.

2.5 Risk Evaluation and Treatment

The manufacturer carries out the following activities to meet the guidance objectives:


System Manufacturer

- review each risk and either avoid, accept, transfer, treat;
- establish methods for relevant risks; and
- implement and verify treatment methods.

The manufacturer decides to review each risk and either avoid, accept, transfer, or treat the risk. It decides to accept the **DSAL0** risk, but as it has determined there is some, albeit low, risk (**DSAL1**) associated with its use of data at a specific point of its lifecycle, it decides to treat that risk. The manufacturer evaluates this risk and considers that the risks should be reduced further by taking some reasonably practicable steps.

Having decided that further risk reduction is necessary, the manufacturer needs to establish treatment methods for relevant risks that are appropriate for DSAL1 data, and in doing so demonstrate that reasonably practicable steps have been taken to reduce the risk. The manufacturer therefore refers to the tables in the DSG document. The manufacturer then documents in its clinical risk management plan:

- planned compliance with the tables;
- the interpretation for the given method / technique (e.g. depth of checking); and
- justification in the case where a technique is not to be adopted.

For **DSAL1** verification data, the tables show that the following are recommended (R) or highly recommended (HR) where loss of data properties integrity, completeness, consistency, continuity, format, accuracy, resolution, timeliness, availability, fidelity / representation, sequencing, intended destination / usage, Goldilocks, and explainability can give rise to a number of hazards. The extracts below indicate the mitigation techniques identified from the tables which should form the basis of the data safety requirements.

Table 2.1: Worked example: filtered techniques tables

Ref	Technique	R / HR	System Design (extracted from ??)
SD.11	Logging facilities	R	Data processing events are logged to allow support staff to monitor the health of the system and provide diagnostic information.
SD.17	Credibility / reasonability checks	R	Dedicated processing implemented to check that data is within reasonable tolerances and / or logically / semantically consistent (e.g., range checks, date checks, record counts, record sizes, special values - not a number (NaN)).
SD.20	Syntax checks	R	Semantic checking of data values and sequences based on defined rule sets.

Ref	Technique	R / HR	Data Design (extracted from ??)
DD.01	Governance model	R	A governance model is established that defines, e.g., data ownership, processing roles and responsibilities, processing authorizations and permissions.

Ref	Technique	R / HR	Data Design
DD.03	Data flow diagram	HR	To describe the data flow in a diagrammatic form.
DD.04	Data model	HR	To articulate how data is organized.
DD.05	Client sign-off	R	Agreement from the client that the system architecture and design is appropriate for the data considered
DD.08	Data dictionary	HR	A collection of descriptions of the data objects or data items in a data model for the benefit of data users.

Ref	Technique	R / HR	Data Implementation (extracted from ??)
DI.01	Review / inspection	HR	Manual review / inspection of data possibly involving data visualization tools.
DI.03	Ground-truth check	R	Inspection against physical measurements (e.g., lengths, positions, heights) taken in the real world.
DI.04	Auditing	R	A period of comprehensive internal and external testing of the data quality process.
DI.09	Authorization	R	A security model is established to control who is authorized to create, view, edit, delete the data.
DI.11	Defined confidence / Trust Levels	R	Criteria are established to provide an objective measurement of the confidence or trust in a given dataset.

Ref	Technique	R / HR	Data Migration (extracted from ??)
No relevant techniques			

Ref	Technique	R / HR	Data Checking (extracted from ??)
No relevant techniques			

Ref	Technique	R / HR	Test Data (extracted from ??)
TD.01	Using informal / ad-hoc means	R	Data is generated by simple means (e.g, spreadsheets, scripts, basic assumptions). There is no formal checking or review of the method of generation.
TD.05	Using manual means	R	Simple test data can be produced by manual means, although this may be prone to human error.

Ref	Technique	R / HR	Test Data
TD.08	Using initial runs of new system	R	This method is often used where the system is breaking new ground and there is no prototype or legacy system to produce test data. Initial operations may differ from eventual usage, so test data must evolve.
TD.09	Derived from real data	R	Where real data is available this is usually a good basis for generating test data (e.g., by modification to increase the test space coverage).
TD.11	Produced by client	R	Ideally the client is involved in producing or at least checking the test data.
TD.12	Client sign-off	R	Where possible, the client should formally agree and sign off the test data as appropriate.
TD.13	Error seeding	R	This is where errors are deliberately inserted into the dataset to demonstrate the effectiveness of data validation.
TD.14	Data reuse	R	Reusing data for one project that was created and thoroughly assured for another project. This can be effective but the read-across should be established.
TD.15	Feedback testing	R	To check output data by comparing it with the input source.

Ref	Technique	R / HR	Media – paper (extracted from ??)
MP.01	Photographic copies	R	Photocopy and store separately.
MP.02	Scan to electronic format	R	Retain both paper and electronic copies.
MP.10	Indexing / cataloguing	R	To support efficient availability.

Ref	Technique	R / HR	Media – electronic (extracted from ??)
ME.01	Regular refresh / rewrite	R	Of magnetic media or flash memory.
ME.02	Suitable physical environment	R	Store media in a clean, low-humidity environment at a steady temperature, cool but not cold.
ME.03	Copies at different locations	R	Physically separate to cover natural disasters, accidental or malicious damage.

Ref	Technique	R/HR	Media – electronic
ME.04	Backups / duplication	R	Backups are essential. Frequency of backup depends on rate of change. The number of generations to keep relates to the impact of data loss.
ME.05	Sample restores	R	Sample restores should be performed at intervals to ensure that the backups are readable and retrievable.

From these tables the manufacturer decides on a series of activities to implement the recommendations that are applicable to its particular endeavour. These activities are expressed as a series of requirements that can be placed on the manufacturer's delivery organization and tracked through to completion.

Table 2.2: Worked example: derived data safety requirements

Ref	Requirement	Guidance Reference
R1	The verification data shall be carefully controlled in the manufacturer's configuration management system. There shall be a configuration management plan that shall define who has responsibility for the data and who is authorized to create and amend it.	DD.01, DI.09
R2	The verification data shall be held on an industry standard file share that is regularly backed up with copies moved periodically to off-site storage. The Backup / Recovery plans shall include periodic sampling of restores.	ME.01, ME.02, ME.03, ME.04, ME.05
R3	The data shall be modelled as a series of patient "journeys" that cover the entire lifecycle of data from first encounter through to archival and deletion of data. The complete set of journeys shall be chosen to exercise all the functionality of the system. The modelling shall include a data dictionary, data flow diagrams and a data model.	DD.03, DD.04, DD.08
R4	To model data from external systems, the manufacturer shall use manual data entry and spreadsheet based records to hold the data.	TD.01, TD.05
R5	The manufacturer has a set of clinical standing data that was used for another system and derived from real data. It includes encounter codes, clinical terms, consultant names, surgery and hospital addresses etc. and can be reused for this system. The manufacturer's Clinical Safety Officer has reviewed the data and agreed its suitability for reuse.	TD.09, TD.14
R6	Some of the verification datasets shall include errors deliberately inserted to check the effectiveness of data validation.	TD.13
R7	The controlled verification dataset shall be subject to review and analysis against defined confidence / trust criteria. Scripts shall be written to check for syntax and semantic consistency of the data and provide a basic credibility check. The scripts themselves shall be validated and verified before use.	SD.17, SD.20, DI.01, DI.03, DI.11
R8	The project shall be subject to an internal delivery quality assurance audit.	DI.04
R9	Data loaded from external system into the system and displayed to the user shall be crosschecked against the original source data, using manual spot-checks.	TD.15
R10	The level of rigour employed in verifying all the above requirements shall be commensurate with the DSAL criticality and so an ISO9001 compliant quality management system shall be adopted.	All

Continued on next page

Table 2.2: Worked example: derived (continued)

Ref	Requirement	Guidance Reference
R11	Data processing events shall be logged to allow monitoring of the health of the system, provide diagnostic information, allow audit trails and support the production of explanations.	SD.11
R12	Each contracted client will be asked to ensure and signoff that test, configuration and adaptation data is appropriate (eg. in terms coverage and representative of real-life values) for their particular circumstance.	DD.05, TD.11, TD.12

The following guidance recommendations were not adopted by the manufacturer for the reasons given. Note that some may however become relevant in the future so actions are set, where appropriate, to review the applicability of the recommendation when the given condition is met.

Table 2.3: Worked example: rejected data safety requirements

Ref	Guidance Reference	Justification	Action
E1	TD.08	The data will be used before any initial run of the system.	Review when data from initial runs is available.
E2	MP.01, MP.02, MP.10	There are no paper based resources for this system.	No further action.

Having determined the requirements arising from the data safety analysis the manufacturer ensures these are included along with other system requirements as part of the overall delivery and operation of the system. It then remains for the manufacturer to implement and verify methods, that is, as well as defining requirements the clinical risk management plan needs to ensure activities are in place to verify and evidence that treatments have actually been implemented.

Likewise, the health organization has identified DSAL2 data and in deciding to treat the risk, it aims to ensure risks are reduced as low as reasonably practicable.

This page is intentionally blank

Chapter 3

Lifecycle Considerations (Discursive)

> *Failure is an amazing data point that tells you which direction not to go.*
> ***Payal Kadakia***

This page is intentionally blank

3.1 Usage Scenarios

If safety-related data is incorrect it can become dangerous when used, either by making a computer or control system perform incorrect actions, or by misleading human users into making incorrect decisions. Since the danger can only be determined when the usage of the data is understood, risk assessment should involve both the consumer of the data and the producer.

Figure 3.1: Consumer-focused integrity requirements

The consumer assesses the use of the safety-related data. (In later phases of the data safety management process this information is used to define the required data properties: for example, how accurate a particular safety-related data artefact must be.)

The producer investigates how the safety-related data is collected and what errors might occur. (Building on activities in later phases of the data safety management process, the producer can provide some form of guarantee, or level of confidence, that the safety-related data meets the specific data-related requirements.)

In some cases a producer will be providing safety-related data without any knowledge of a specific user (e.g., mapping data or generic databases that are sold to many users). In these cases the producer will need to make some assumptions about possible users, and then clearly state what level of integrity the data has been produced to. It is then up to the users to check whether the declared integrity matches their need.

This page is intentionally blank

3.2 Data in System Lifecycles

Like other components of a safety-related system, the safety dependency of data is dictated by the context in which it is used and the causal links that become established where loss of one or more of the required properties can contribute to hazardous system states. For example, a given dataset (say configuration data) could be used in a number of separate contexts such as:

- prototyping a system to demonstrate solution feasibility of a safety-related system;
- development testing of a safety-related system; and
- live operational use of a safety-related system.

In these cases, the dataset is the same but the context of its use changes the safety significance and therefore the level of assurance that it may require. It follows that the DSAL of a dataset is also predicated on where and when in the lifecycle the dataset will be applied.

To illustrate this concept, a number of generic model lifecycles are discussed below. Note that these are not intended to be prescriptive or mandate the use of any particular model. Instead, they are being used to illustrate how the Data Safety Management Plan could articulate these lifecycle considerations.

Development: the diagram in Figure 3.2 represents a typical development lifecycle using an iterative development approach[1]. In this model there are key phases as the system transitions from concept through to testable executable code. The process is iterative in that several cycles of functional elaboration, design, development and test may be run and these typically will focus on the areas of the system that bear most technical risk or comprise the key functional use cases so the client gets early visibility of the system. This early awareness allows feedback to be provided into the next iteration to help steer the solution to the client's actual needs. Traditional waterfall implementation can map onto this model on the basis that there is only one iteration in each phase and all activities in one phase need to be completed before progressing to the next.

The model itself may vary depending on the specific needs of the project but the diagram illustrates that different data categories become significant at different points of the process.

Operational Once a system has been developed it will move into an operational lifecycle or indeed, if data safety has not previously been considered for an enterprise, then the system could already be in operational use. These operational lifecycles tend to be cyclical in nature; the diagram in Figure 3.3[2] illustrates a typical model.

Again, specific data will come into play at different periods in the process. Documenting the relationship between process steps and data categories will therefore give clarity as to when a particular assurance technique needs to be applied.

Data supply chains The previous models relate to typical system supply and operate perspectives but there are also other data supply chains where a number of organizations engage in the procurement and use of safety-related data. These processes may include the development and operational lifecycles but a different model is required to fully represent the wider processes that are being employed. The diagram in Figure 3.4 shows such a model representing a data acquisition lifecycle.

This model represents the interactions between three key organizations:

[1] The diagram is based on International Business Machines Corporation (IBM)'s Rational Unified Process, an iterative software development process framework. The original diagram is in the public domain.

[2] ITIL is a registered Trade Mark of AXELOS Limited. All rights reserved.

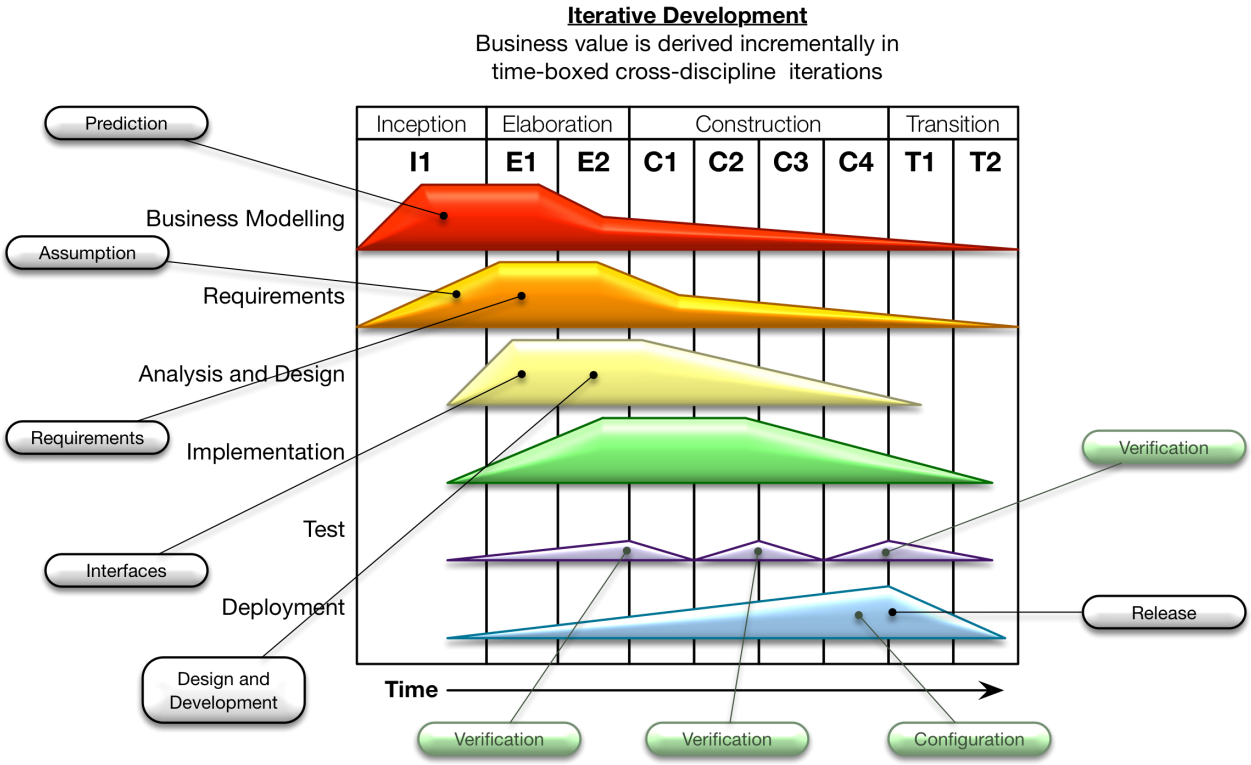


Figure 3.2: Development lifecycle

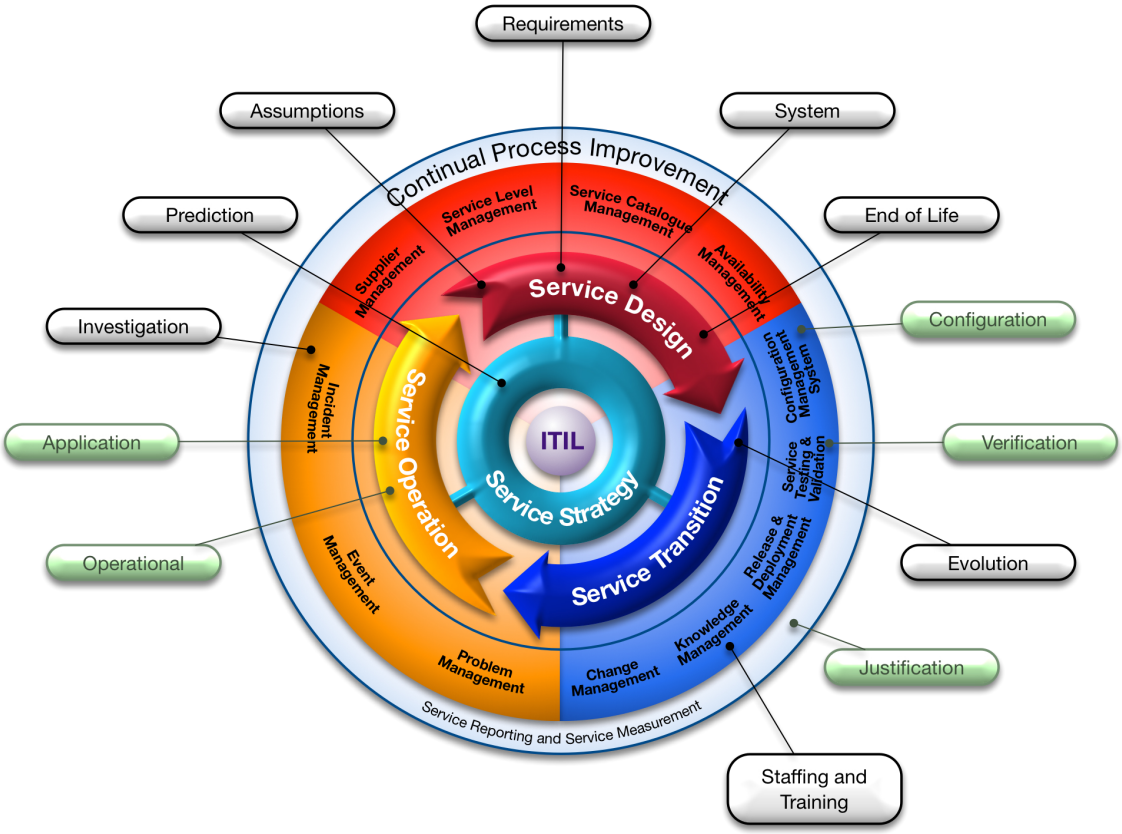


Figure 3.3: Operational lifecycle

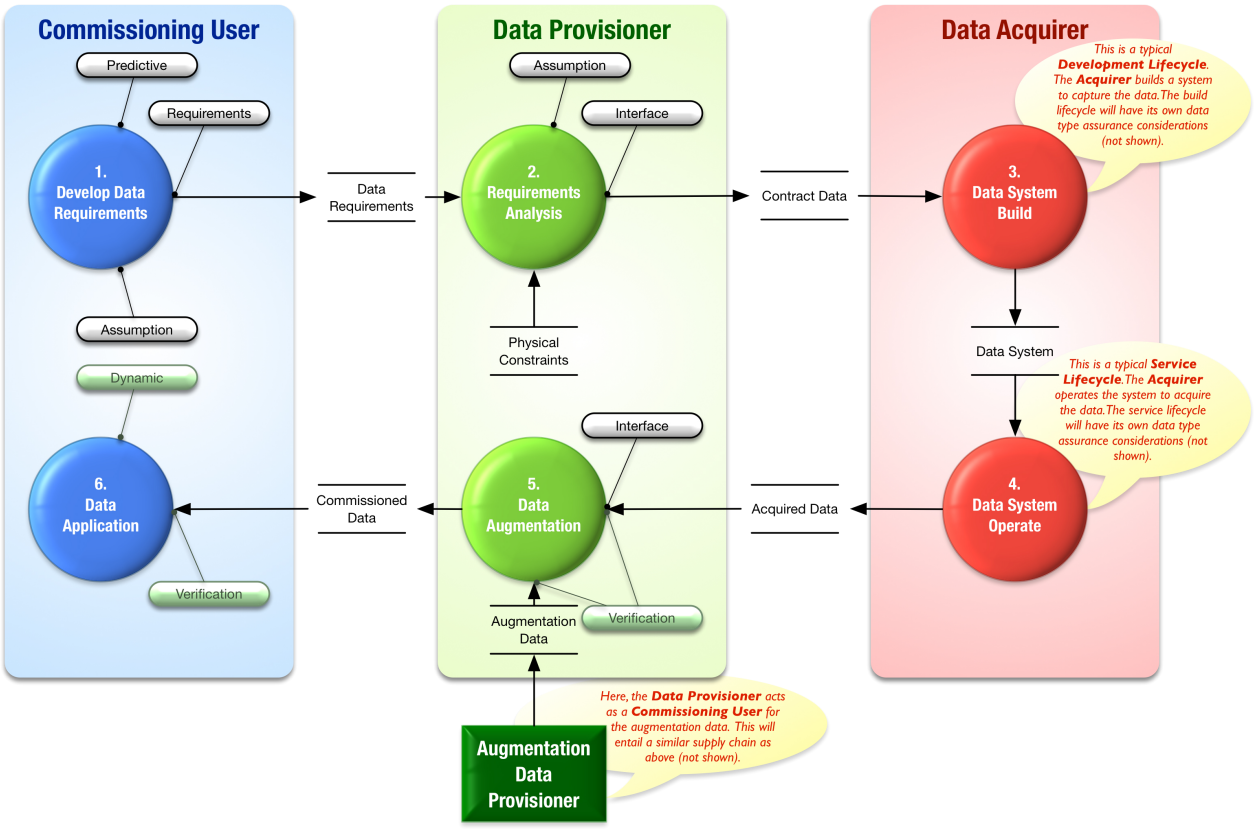


Figure 3.4: Data supply chain

- The commissioning user: the organization that has the need for the data;
- The data provisioner: the organization that will fulfil that need for data; and
- The data acquirer: the organization employed by the Data Provisioner to carry out physical collection of data.

Note that these may be three separate organizations, or they may be separate business units within the same, larger, organization.

In this supply chain, the commissioning user is a consumer of the data and the data acquirer is a producer of data. The data provisioner acts both as a consumer (from the data acquirer) and producer (to the commissioning user) of data. Similarly, an organization that augments datasets is both a consumer and producer of data in the supply chain.

The commissioning user requirement analysis is the key process step where the commissioning user's expectations for data are agreed with the data provisioner. The requirements may be adjusted because of physical constraints (e.g., loss of precision because of physical measuring device constraints) and may include additional requirements to augment the captured data with additional information (e.g., airport codes added to a measurement of a given runway length).

The data provisioner may employ a data acquirer to capture the data (e.g., to carry out a physical survey of a site). The acquisition phase may itself require a specialised system to be built to perform the capture and data refinement to meet the data provisioner's specifications. Such systems will then themselves be subject to the development lifecycle model considerations discussed above. Likewise, the data augmentation phase may require further system development processes or indeed, could trigger an instance of the model

again as the data provisioner acting as a Commissioning User.

Acquired and augmented data is then fed into the operational system that has been built for providing the service of generating the commissioned data. This system in its service provision role would then typically follow the operational lifecycle process model discussed earlier.

3.2.1 Tool Assurance

Tools in this context are considered anything that automates all or part of a process, for example, data creation or data transformation. Test tools are also included (i.e., the term is not limited to parts of an operational system).

Tools can impact data safety in different ways, depending on both their function and how they are to be used. For tools to be considered fit for purpose it is necessary to show that the tool meets its requirements in the context in which it is to be used. The activity to ensure a tool is fit for purpose is usually called "tool qualification".

The first step is to define the purpose for which the tool is required to be fit. Once that is done, and the tool's requirements are specified, there are three main strategies available for qualification:

- Use evidence of a previous certification of the tool by a trusted third party (unlikely to be available in most industry sectors);
- Base tool qualification on the practices used when designing and developing the tool (only practicable for tools developed within the organization); and
- Use one of the available industry-specific guidance documents that admit COTS solutions, e.g., EUROCAE Document ED-215 (RTCA/DO-330) [**?**].

Further details on tool qualification are presented in Appendix **??**.

3.2.2 Test Data

The generation of suitable test data is critical to verification of a safety system. The test data must include both representative "normal" values based on intended usage and also values which push at, and beyond, normal use to provoke hazards that the system might produce. This latter type of test data is particularly hard to generate; generally it must be credible, yet it must stress the system to react in a way that the preservation of safety properties can be assessed.

In general, all the properties of the test data should be considered and an assessment made as to whether breaking a property (e.g., introducing corrupt or late data) would cause a problem to the system. If it does, then specific test data should be produced to facilitate testing of this potential problem.

Some suggestions for test data for safety-related systems are:

- Use of values on or around boundaries;
- Use of extreme values, way beyond what could be reasonably expected;
- Use of typical "everyday" values / sets;

- Some realistic but unexpected values;
- Try combinations of data values or data items that are problematic together (e.g., inconsistent);
- If possible, use some values known to have caused problems in the past;
- Where appropriate, use values related to timing, rollover or date boundaries;
- Where possible, use white box values (i.e., those derived from an understanding of the system);
- Use a set of values with drift or bias over time;
- Use datasets with particular statistical properties (e.g., distribution, patterns etc.);
- Use data which has incorrect formatting, ordering, or out of sequence, etc.; and
- Try data which has repeated sets of values or pseudo-random characteristics.

Typically very complex test data is derived from recorded live feeds of real data flows. While this data can be extremely useful for regression purposes, it should be recognised that it is unlikely to contain many outlying or boundary data items. Therefore it may need to be modified to test any hazardous situations; this modification can be difficult and may require sophisticated tools to both ensure correct properties and injection of the intended faults (for instance to introduce a statistical bias to the data).

Simulator / emulator derived values can be useful, but again the issue is how realistic the values are: often the accuracy, resolution or timing of simulated values may be different to real data.

Coverage with test data is something to consider. Sometimes the same dataset is used for multiple test scenarios, when in fact it is not stressing all of them to the same degree. Test data coverage can be collected over requirements, code or design, but it is important not to forget hazards: coverage of the hazards and mitigations identified in the hazard log is a key aim.

In general some measure of the quality and suitability of the test data can be useful. This could be based on statistical properties, coverage of hazards or coverage of requirements.

Test data must show continued relevance, through systems evolution and over time. It is good practice to build up extensive regression suites containing coverage of all detected problems to date.

3.2.3 Interfaces with Existing Assessments

3.2.3.1Data and Software

Although most people feel they have an intuitive understanding of the difference between software and data, upon closer examination the boundary is not always as clear as it may first appear.

Consider, for example, Java bytecode, which is operated on by a Java virtual machine. From one perspective, it could be argued that the Java bytecode is simply data. By extension, it could also be argued that the Java source code is also just data. This type of argument can be extended to suggest that any software can, at least from one viewpoint, be considered as data. Conversely, think about the data used in a 3D printer, perhaps to produce a part for an aircraft. This data could be viewed as a program for the printer; that is, it could potentially be viewed as software. This type of argument can easily be extended across a range of situations, especially those relating to configuration data.

While they are interesting, and potentially important, these philosophical considerations should not detract from the practical issue: there are some aspects of data (using the term in a generic sense) that are often not explicitly addressed in standards. These are a consequence of features that are more readily apparent in data than in software. Examples include:

- It is easy for data to be reused in a range of contexts and despite appearances it is not trivial to translate an assurance argument that the data is fit for purpose from one context to another.
- It is not always clear who owns or is responsible for data, especially when data is shared and processed amongst a collection of disparate systems.
- Data often has a lifetime, that is a time after which it is no longer valid. This may be a strict cut off, or a more gradual degradation in the utility (or applicability) of the data.
- There is often a default value for data. While this can make systems easier to use and hence more productive it can be difficult to identify a default value that is appropriate for all circumstances.
- It can be easy to change data. In some circumstances this can give rise to a temptation to make uncontrolled and potentially untested changes. It can also allow data to be fraudulently changed after an accident.

In summary, data and software are closely related and, as such, need to be considered together in system engineering activities, including system safety analyses. However, data and software emphasise different facets of risk and they are susceptible to different mitigation approaches; this means there is also a need to adopt a data-focused perspective. It also means that software assurance levels cannot be mapped directly to DSALs.

3.2.3.2Data Safety and Security

When generating high-level processes and techniques to manage the risks posed by data, it is worthwhile understanding the difference between the safety risks posed by accidental failure to preserve Data Properties and the security risks posed by actors maliciously undermining the properties of data.

The relationship between safety and security, as engineering concepts, can be summarised by their relationships to cultural, developmental and aspirational properties of systems development.

Culturally, embedding both safety and security into an organization is seen as a key strategic goal for creating systems that are both safe and secure. Developmentally, safety and security are quality factors, generating transverse requirements that impact the entire system. Most importantly, at the aspirational level, both safety and security have the common goal of preventing harm from accidental and malicious interventions respectively.

For an organization aiming to create systems that are both safe and secure, these connections can be both a benefit and a burden. The shared goal of preventing harm means that both quality factors seek to identify routes to harm through analysis of the system being developed. This can result in shared processes and tools, which in turn can save time and money during systems development. However, safety and security interact in a more volatile way at the functional level. Security failings can undermine the safety case for a system and, conversely, safety requirements can prevent the implementation of standard security solutions. For example, the German government published a report in 2014 into a fire at a steel works caused by a cyber attack that resulted in the control system being placed into an unsafe state and the

safety system being unable to intervene (Section 3.3.1 of [**?**] - in German). In addition, "fail-safe" states can often leave a system with exposed security vulnerabilities.

These links between safety and security infer that there are connections between the sub-categories of data safety and information security: both attempt to take a data-centric view of the system of interest in order to improve the associated quality factor; and both attempt to prevent harm through the preservation of the properties of data within that system.

In the security domain, the three key properties of data considered are confidentiality, integrity, and availability. Confidentiality, the failure of which is termed "information disclosure" in the Microsoft Security Model, [**?**] is typically not a safety concern as, without malicious intent, information sharing is not inherently unsafe. However, when considering systems where confidentiality is an important property, the interaction between data safety and security cannot be trivially resolved. For example, accidental disclosure of information can form part of a causal chain which leads to harm from a malicious actor.

Data integrity is a critical property for both domains. The Microsoft Security Model describes malicious removal of the property of integrity as "tampering". Whether by accident or through malicious intent, the potential harm from loss of data integrity can be disastrous to a safety-critical system, from the values of drug dosages to control system parameters.

Data availability is also important to both domains. Loss of availability, or "denial of service" in the Microsoft Security Model, is another property that can be lost accidentally or through malicious intervention. Loss of availability prevents systems from functioning properly and can result in undefined behaviour if not mitigated by design.

Further guidance on the integration of safety and security can be found in a code of practice published by the IET [**?**]. The Code of Practice is written for engineers and engineering management to support their understanding of the issues involved in ensuring that the safety responsibilities of an organization are addressed, in the presence of a threat of cyber attack.

This page is intentionally blank

Chapter 4

AI and Machine Learning (Informative)

> *Learning never exhausts the mind.*
> ***Leonardo da Vinci***

This page is intentionally blank

4.1 Machine Learning Training Data

A complete discussion about the assurance of ML /AI algorithms is beyond the scope of this document. However, as such algorithms are inherently data-driven, the data safety guidance provides the following notes on good practice for datasets used for ML / AI applications.

The data for real-world ML scenarios may be effectively infinite (for example, with self-driving cars, where the environment is largely unbounded) and can only be approximated by a finite number of scenarios. Therefore it is crucial to have an extension process in place to extend the system when previously unknown situations are experienced. ISO 21448 [**?**] captures this process by introducing the term safety of the intended functionality (SOTIF) and requires the systematic handling of unknowns. Although specifically aimed at road vehicles, SOTIF may be of interest to other domains.

For more general guidance on managing the safety of autonomous systems, please refer to the safety assurance objectives guidance produced by the SCSC Safety of Autonomous Systems Working Group (SASWG) [**?**].

This page is intentionally blank

4.2 Data Categories

Three data categories are discussed: ML training data (the data used to train the model), ML test data (used to provide a measure the trained model's accuracy), and ML validation data (used to independently verify the trained model).

This page is intentionally blank

4.3 Real-world Data

It is important to obtain sufficient quantities of data for the training data. It is generally assumed that this is sufficiently representative of situations that could be encountered to allow the algorithm to generalise the appropriate behaviour for all encountered situations. However, there are problems with real-life training data. Firstly, it is possible that rare, hazardous cases are never encountered during the capture of a dataset (as those events have been engineered to be statistically unlikely) and are too dangerous, or expensive, to generate deliberately. Poor representation of rare cases in training data will result in a model that is insufficiently prepared to handle those important scenarios. Poor representation (or weighting) of rare cases in ML test data or ML validation data can result in error heuristics that do not provide adequate insight into the system's performance in rare situations. A second issue with real-world data is that amassing, understanding, verifying and validating it at scale is resource intensive. Hence it is not just a case of more testing, a more cost-effective approach is needed. Some possible techniques are: artificial ML test dataset generation, data fault injection and simulation. It may well be more practicable to use simulated training data during the training phase to give more control and avoid the cost and effort of real-world testing. The issue then is how realistic the simulated data (and the associated simulation) can be. Note that artificial datasets can also be used to increase the frequency of rare events, thus biasing the algorithm to managing those rare events effectively, but doing so, could impair the "sunny day" performance of the system.

This page is intentionally blank

4.4 Sensor Data

In many cases sensors will be on-board and embedded (e.g. on an automated vehicle) and will be subject to many real-world influences. Sensor data may be changed or degraded by both internal influences, e.g. system environment temperature, and external factors e.g. humidity, ambient light, weather conditions such as snow or fog, and rarely encountered extremes such as lightning or flood. Sensors may also suffer degradation due to ageing and specific faults due to their positioning, construction or configuration. Sensors may need to be diverse to increase resilience and give an overall more reliable output. Realistic data variations, informed by an understanding of likely sensor failure modes, need to be incorporated in the training dataset (c.f. the AoA sensor in the Boeing 737 MAX 8, discussed in **??**). Sensor fusion can effectively create "composite data", e.g. from merging of lidar, radar and camera data; this will have different (and possibly hard to predict) properties compared to the individual sensors.

This page is intentionally blank

4.5 Data Coverage

Some notion of data coverage, that is, how well the training data covers the intended operational domain, is needed: this could be based on use of all known datasets, all known hazardous situations or some other measure. White box techniques involved in exercising all connected sensors or coverage of nodes in a neural network may have some value. Regardless, it will be essential that the training data covers all required safety test scenarios.

This page is intentionally blank

4.6 Dataset Assurance

There is a need to have assurance about the datasets themselves, including their configuration management. Assessment of the quality and suitability of the training datasets should be included in the safety argument.

The University of York has developed a method for the assurance of machine learning for use in autonomous systems (AMLAS) [**?**]. The University's approach is intended to assist in the development of a compelling argument about a machine learning model, to feed into a system safety case.

This page is intentionally blank

4.7 Data Diversity

The degree of data diversity is important. Common data items across datasets need to be avoided, especially across training and ML validation data. Common data may include data derived from the same sources, environments, simulators, tools, techniques or mathematical models.

This page is intentionally blank

4.8 Data Properties

Bias, sufficiency and accuracy are all concerns with the data. fidelity / representation may also be a problem as modelling real-life sensor data is difficult. Errors should be introduced for training: this may be by analytical means, by random, automated sampling of data to determine the degree of similarity between samples, or by deliberate insertion of known edge cases. It is suggested that all the data properties mentioned in this guidance document are considered.

This page is intentionally blank

4.9 Hallucinations

The human brain can see and interpret patterns in unexpected ways, for example seeing a shadow that suggests a monster is waiting to pounce, or seeing an animal in a cloud formation[1]. AI tools can similarly interpret data in novel and unplanned ways, perhaps seeing patterns that a human would not perceive, or seeing shapes in noise. However with AI tools, hallucinations can appear in two different ways – in the interpretation of the question or instruction to perform, as well as in the interpretation of the source data. The problem is exacerbated in the case of AI "chatbots", where the output presentation in natural language can make the reply seem authoritative and result in it being trusted, as though the AI chatbot is an expert in the field under consideration [**?**].

The complete eradication of hallucinations may not be possible, so it can be dangerous to rely fully on the output of an AI without verification. Biases in the training data are regarded as a significant cause of hallucinations, so selection of training data should include consideration of its quality – whether it is diverse, balanced and well structured [**?**]. In addition, evaluation of the resultant behaviour should be assessed by humans, to achieve "reinforcement learning with human feedback" [**?**]. The results of the assessments can then be fed back into the model, thus fine-tuning it for its specific application.

Note that fine-tuning will generally only be possible for a specific application and / or a specific environment. As with all safety analyses, it is critical that these be clearly defined.

We asked two different AIs to comment on the risks that hallucinations present to data safety. Their replies were as presented in the following two boxes:

Response of the AI

AI hallucinations can have significant implications for data safety and overall trust in AI systems. Here are some key concerns:

Misinformation and Disinformation: If an AI model generates incorrect or misleading information (hallucinates), and this information is used without verification, it can lead to the spread of misinformation. In sensitive areas like healthcare, finance, or law, this could have serious consequences.

Decision-Making Risks: In scenarios where AI is used to aid decision-making, hallucinations can lead to incorrect or risky decisions. For instance, if a financial AI advises on investments based on hallucinated data, it could lead to substantial financial losses.

Trust and Reliability Issues: Frequent hallucinations can undermine the trust in AI systems. Users and organizations may become reluctant to rely on AI for critical tasks if there's a significant risk of the AI generating false information.

Security Vulnerabilities: AI hallucinations can be exploited in cybersecurity. For example, if an AI system responsible for network security misinterprets data or hallucinates a threat, it could either overlook actual threats or allocate resources to non-existent threats.

Privacy Concerns: There's a risk that an AI model might inadvertently reveal or imply private information in its outputs, especially if it has been trained on sensitive datasets. While not a hallucination in the traditional sense, this is related to the broader issue of how AI handles and interprets data.

[1] These are not strictly hallucinations in the medical sense, but serve to illustrate the issue with AI interpretation of data.

Ethical and Legal Implications: AI hallucinations can lead to ethical and legal challenges, especially if they result in harm or damage. Determining liability in cases where AI-generated misinformation leads to real-world consequences is complex.

Bias and Fairness: Hallucinations can also be a reflection of biases in the training data. An AI might generate outputs that are unfairly biased against certain groups or individuals, leading to ethical concerns and potential legal issues.

To mitigate these risks, it's crucial to have robust validation and verification processes, use AI in conjunction with human oversight, continually update and improve AI models with diverse and accurate data, and maintain transparency about the limitations and capabilities of AI systems.

Response of the AI

The implications of hallucinations for data safety are complex and far-reaching, spanning aspects of privacy, security, and misinformation. Here's a breakdown of some key concerns:

Privacy: False generation of personal data: AI hallucinations can fabricate personal information (names, addresses, financial details) that doesn't actually exist, potentially violating privacy rights and harming individuals. Exposure of sensitive data: If trained on sensitive data, hallucinations might leak this information unintentionally, even if not explicitly instructed to do so. This could include medical records, financial data, or other confidential information. Security:

Creating vulnerabilities: Hallucinations might invent security flaws or exploits that don't actually exist, leading to wasted resources and potentially even vulnerabilities being introduced into real systems. Misleading security assessments: If used for security testing, hallucinations could create false positives or negatives, providing inaccurate assessments of a system's security posture. Misinformation:

Spreading false information: AI-generated hallucinations can easily spread misinformation, as they appear convincing and authoritative even if factually incorrect. This can have significant implications for public discourse, trust in institutions, and even manipulation. Evolving deepfakes: Hallucinations could be used to create even more sophisticated deepfakes, making it harder to distinguish between real and fake content, further eroding trust and potentially influencing elections or other high-stakes situations. Additional considerations:

Impact on trust: As reliance on AI grows, hallucinations can erode trust in its outputs, making it harder to distinguish reliable information from fabricated content. Regulation and accountability: Establishing clear guidelines and accountability mechanisms for AI systems prone to hallucinations is crucial to mitigate their risks. Overall, addressing the implications of hallucinations for data safety requires a multi-pronged approach:

Developing robust and transparent AI systems: Techniques like adversarial training and explainable AI can help reduce hallucinations and improve transparency. Implementing data governance and privacy measures: Strong data protection frameworks and responsible data handling practices are essential to minimize the risk of sensitive data exposure. Promoting media literacy and critical thinking: Educating users to critically evaluate information, regardless of its source, is crucial in the age of AI-generated content. Establishing ethical guidelines and regulations: Clear ethical frameworks and regulations for AI development and deployment are necessary to ensure responsible use and mitigate potential risks. By acknowledging the dangers of hallucinations and taking proactive steps to address them, we can minimize their impact on data safety and build trust in AI-powered technologies.

4.10 Dataset Management

Datasets will be large and complex and ensuring quality and consistency will be difficult. Measures will have to be taken to control, manage and archive large amounts of data. Datasets should be subject to formal change control and monitoring. Tools will be needed to manage the data.

This page is intentionally blank

4.11 Optimization

What is an optimal size of a training or ML validation dataset? A dataset may contain many repeated or redundant values. If these do not add to the training outcome it is possible that some values may be removed. Also if there are ways in which the dataset could be compressed or optimized then these should be considered to assist management and curation.

This page is intentionally blank

4.12 Compliance

A “safety by compliance” strategy for such data is difficult to implement as there are currently few existing standards and little guidance for data in this area. It is therefore difficult to quantify data sufficiency and as a result, no generally agreed “ALARP” principle for data.

It may be possible to demonstrate Data Safety through statistical methods, such as when millions of representative miles have been driven. However such approaches have generally been found insufficiently robust to support safety-critical systems in other domains such as aviation, and so must be treated with caution. The demonstration of compliance for the data is likely to require a compliance argument for the tools that produced the data, the preservation of data properties as described within this document, and an update process such as that specified in ISO 21448 (SOTIF) [**?**].

This page is intentionally blank

4.13 References

This appendix provides only a brief overview of the issues around the application of data safety techniques to machine learning. It is based upon information in the following documents, which should be consulted for further details:

- Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges [**?**]
- Quantifying Data Set Properties for Systematic Artificial Neural Network Classifier Verification [**?**]
- Safety Critical Integrity Assurance in Large Data Sets [**?**]

This page is intentionally blank

Chapter 5

AI and autonomy (Informative)

> *The potential benefits of artificial intelligence are huge, so are the dangers.*
> ***Dave Waters***

This page is intentionally blank

5.1 Job Displacement

AI and automation can lead to the displacement of jobs, as machines can perform some tasks more efficiently than humans. This can have significant socioeconomic implications, including increased unemployment and wage stagnation in affected industries. The International Monetary Fund (IMF), in January 2024, gave an estimate that AI will impact 40% of jobs globally, and 60% of highly skilled work, ([**?**]). The report balances the already known complementary nature of AI to human work against the likely detrimental aspects. Unusually for technological advances, AI is expected to impact high-skilled jobs to a greater extent than manual skills. The report states, "advanced economies face greater risks from AI – but also more opportunities to leverage its benefits". Labour requirements could significantly reduce, wages could be lower and of course jobs will disappear. Managing director of the IMF, Kristalina Georgieva, said that "AI will likely worsen overall inequality , a troubling trend that policymakers must proactively address to prevent the technology from further stoking social tensions." ([**?**]). How economists are changing the standard growth view of economics and how engineering and other businesses can engage with such developing problems is discussed in an article in the February 2024 edition of the SCSC Newsletter ([**?**]). The data used to determine the impact of an AI system, especially when it affects currently underdeveloped nations, must ensure that the social tension Georgieva refers to does not reach a breaking point. Like well digging, rail construction and careful dam placement, AI technology could raise many millions out of poverty if it is applied to benefit life on earth.

This page is intentionally blank

5.2 Bias and Discrimination

AI systems can inherit biases present in their training data, leading to discriminatory practices. For instance, facial recognition software has been shown to have higher error rates for people of certain racial and ethnic groups. At a very basic level, slight imperfections in coatings for car registration plates, easily ignored by humans, has made some number plates unrecognisable to automatic number plate recognition cameras at car parks. This can be easily missed in training the system since it would have been difficult to anticipate every possible imperfect state of a number plate. In this case advances are made from experience but there are many situations in which day one of a system's deployment has to be near perfect. Training data sets should be vetted by committees, as safety usually is and should be always. That technique is more likely to bring out overlooked areas, obvious to someone on the committee but totally missed by others. Testing of developed systems should also cover the full range of possibilities, whether or not those involved in the development provide assurances, as assurances can hide failings.

This page is intentionally blank

5.3 Loss of Human Skills

With AI taking over tasks such as navigation or memory-dependent activities, there is a risk that humans may lose certain skills that are not used regularly, potentially reducing cognitive abilities over time. However, the opposite is also true because, driven by the threat of automation, many people could turn to artisan skills to create unique or less ubiquitous products providing an element of appealing individuality. This could turn supermarket checkout staff into, for instance, entrepreneurial upholsterers or cosmetic developers when and where all checkouts have become automated. Regardless of whether or not skills are lost, it is certain that in order to verify and validate data intensive artificial intelligence systems someone will be required to maintain those basic skills to provide the assurance. To make a simple analogy, one cannot validate a French grammar correcting AI system if one does not speak French fluently.

This page is intentionally blank

5.4 Security Risks

AI can be used to develop sophisticated cyber-attacks, and AI systems themselves can be vulnerable to such attacks. The integration of AI into critical infrastructure heightens the potential impact of these risks. The issue of autonomous vehicles in military applications has been discussed in an SCSC newsletter, ([**?**]), which quotes "Formal Verification of Ethical Choices in an Autonomous System" ([**?**]): "All participants in society are required to follow specific regulations and laws. An autonomous system cannot be an exception". Given this as an accepted principle, it is then incumbent on engineers to design systems which follow ethical regulations. This presents difficulties and it demands that the most sophisticated sensors be employed on killer robots when decisions will need to be made as to whether or not to engage in an action which may result in death or excessive damage. Such examples are, firing a missile at a market square or bridge where known enemies are observed. At what point does the death of others not engaged in wrong doing become acceptable, known as collateral damage. It is important to stress that the data used to train systems covers all possibilities, just as in human vs. human conflict. If the machine is not programmed to make the decision, the machine must request clarification from human controllers. The issue is that it will almost certainly prove impossible to develop military killer systems that can be programmed with sufficient data to make every decision required in all circumstances without needing to consult humans, though direct accurate hits on isolated targets with little risk to the combatants of the targeting force is likely to remain an acceptable usage. But it would be sad if the governments of the world end up pitting machines they don't understand against each other with all the horrendous possible consequences.

This page is intentionally blank

5.5 Privacy Erosion

AI's ability to analyse vast quantities of personal data can lead to erosion of privacy. For instance, AI can be used to make highly accurate predictions about individuals' behaviours, preferences, and even future actions. To a limited extent this is already operational and desired by police forces wanting to identify faces and even the gait of wanted persons. The usual claim is that "If you are not doing anything wrong you have nothing to fear". But many democracies are turning against the principle of eroding privacy, considering some policing techniques AI makes available are too intrusive. Even when the nation is particularly controlling, the creation of the network of cameras across a nation does provide the environment for a takeover by controlling elements. The Metropolitan Police of London are using live facial recognition (LFR) cameras ([**?**]). Their policy is "LFR cameras are focused on a specific area; when people pass through that area their images are streamed directly to the live facial recognition system. This system contains a watchlist: a list of offenders wanted by the police and/or the courts, or those who pose a risk of harm to themselves or others." They currently assert that "LFR is not a ubiquitous tool that uses lots of closed circuit television (CCTV) cameras from across London to track every person's movements". But a change of national or local government could change that. The risk is that data may be abused or false.

This page is intentionally blank

5.6 Control, Autonomy, Ethical and Moral Considerations

As AI systems become more autonomous, there is a risk that they may act in unforeseen ways that are not aligned with human intentions or may be manipulated to act against human interests. Also there are significant ethical questions around AI, including the morality of decisions made by AI systems, particularly in life-and-death situations such as in autonomous vehicles or military applications. Although mainly concerned with military questions "The Troubling Aspects of Autonomy" ([**?**]) discusses the issues of control that arise for dangerous life-threatening situations and is pertinent to non-military systems.

This page is intentionally blank

5.7 Economic Inequality

The benefits of AI may accrue disproportionately to those who own the technology, potentially exacerbating economic inequality. Companies and nations that can invest in AI could gain significant economic advantages, leaving others behind. The likely beneficiaries are the developed economies. 40% globally will be affected according to an IMF discussion paper and 60% in the advanced technology world. In the fast world of competitive developments in AI which has now begun, it is difficult to believe that the potential pitfalls are ever fully examined when, just as in more conventional engineering systems, the bottom line remains money gained through business success. Only regulation by governments and other authorities can be expected to be effective in limiting damage outside the bounds of AI systems. Some are now advocating a new form of economics known as Doughnut Economics as advocated by Professor Kate Raworth, senior associate at Oxford University's Environmental Change Institute in her book on the subject ([**?**]). A brief discussion of the some of the issues of such economics can be found in ([**?**]). There could be ways to support business in developing countries if AI tools are made available on smart phones at low cost to, for instance, African women's business collectives. Such a system may just have emerged with China's DeepSeek tool. Data input and received from such applications will need to be safe and accurate if it is to benefit the less well-off of the world. Businesses must have good data to when using AI that provides reasonably accurate appraisals of outcomes from various actions. Ultimately, nobody should be left falling short of the social foundation that the doughnut economy declares the future 'bottom line' to be. In the appendix to Raworth's book, a table is given showing the issues that need addressing and declaring the sources. Such sources should be consulted for trusted data when new projects are developed.

This page is intentionally blank

5.8 Dependency

Over-reliance on AI can lead to a lack of preparedness when systems fail. If critical infrastructure or services are AI-dependent, outages or malfunctions could have severe consequences. Simple growth of single point failure understanding applied to AI should mitigate a lot of the risk of loss of urgent data, such as fly-by-wire aircraft having three computers with a voting sub-system in overall control.

This page is intentionally blank

5.9 Manipulation and Fake Content

AI can be used to create deep-fakes and synthetic media, which can be used to manipulate public opinion, perpetrate fraud, or spread misinformation. The Partnership on AI (PAI), ([**?**]), has pursued the concept of there being a spectrum of harm from manipulated media. How this will all resolve is as yet unknown; the problem is in its relative infancy. With partners, a working definition arrived at is "any image or video with content or context edited along the 'cheap fake' to 'deepfake' spectrum with the potential to mislead and cause harm" ([**?**]). A diagram and explanation of the spectrum can be found at Data and Society's website ([**?**]). Twelve principles for labelling manipulated data can be found at the PAI on AI ([**?**]) . As to whether any of this is enough, only time will tell. Much is seen in the news of demands for content providers to do better at eliminating bad data, be it manipulated or such things as sites encouraging suicide. However, it is worth noting that The Future of Humanity Institute (FHI), a research group within the Oxford Martin School, has joined the PAI and that organisation has as members many of the known big players in the data intensive media field: Amazon, Apple, Google/DeepMind, Facebook, IBM and Microsoft. There goal is to formulate socially beneficial best practices for AI development. How much their belief in their achievements coincides with politicians requirements we must wait and see. The PAIhas produced a document "PAI's responsible Practices for Synthetic Media", ([**?**]), which argues the best way to overcome fake media is by the organisations not trying to deceive anybody to have codes of practice. They seek to advance ethical and responsible behaviour. Synthetic media, also known as generative media is defined as "..visual, auditory or multimodal content that has been generated or modified, commonly by artificial intelligence."

Legitimate use of such media may be, for example, entertainment, art, satire, education, and research. Furthermore, techniques can be used legitimately or harmfully; there is no barrier. One recommendation for technology and infrastructure creators is "aim to disclose in a manner that mitigates speculation about content, strives towards resilience to manipulation or forgery, is accurately applied and also when necessary, communicates uncertainty without furthering speculation". The guide goes on to recommend practices for 'Creators' and 'Distributors and Publishers'. Transparency and disclosure are the main thrusts. It therefore appears that the legal definition for consumer risk will mirror goods purchase: *caveat emptor*, let the buyer beware, but with safeguards provided by companies wishing to take ethical stances.

This page is intentionally blank

5.10 Existential Risk

In the long term, some theorists and scientists worry that the creation of super-intelligent AI could pose an existential risk to humanity if it becomes uncontrollable or its goals are misaligned with human values. This particular aspect of data safety with AI systems shows how important it is that the data used in AI systems reflects humanity's values. The doughnut economic model uses, as the social foundation for humanity, twelve of the seventeen goals of the United Nations. Any AI system that is not trained, where relevant, using data that will ultimately achieve those goals, is almost certainly not an ethical system designed to align with the established values of the global humanitarian consciousness. Also, many systems capable of destroying that have been granted autonomy, such as weapons, need to have been trained using the relevant data reflecting relevant guidelines (in the military case, on conflict rules). Issues related to the military problems are discussed in [**?**].

This page is intentionally blank

5.11 An AI's view of the risks

We asked an AI "What sort of mistakes due to AI can be hard to detect?." The AI replied as follows:

Response of the AI

Artificial intelligence (AI) systems, while powerful and increasingly integral to many aspects of modern life, are not infallible. There are several types of mistakes or errors in AI systems that can be particularly hard to detect:

Subtle Data Bias: AI algorithms are often only as good as the data they are trained on. If this data contains biases, the AI can perpetuate or even amplify these biases. These biases can be subtle and deeply ingrained in the data, making them difficult to detect without thorough analysis.

Overfitting: This occurs when an AI model is too closely tailored to the specifics of its training data and fails to generalize well to new, unseen data. Detecting overfitting can be challenging because the model may perform exceptionally well on training data but poorly in real-world applications.

Black Box Models: Many advanced AI models, like deep neural networks, are often considered "black boxes" because their decision-making processes are not transparent or interpretable. This lack of transparency makes it difficult to understand why the AI made a certain decision and, consequently, to identify when it makes mistakes.

Contextual Misinterpretation: AI may not fully understand or consider the context in which data exists, leading to misinterpretations. These errors can be hard to detect because the AI's output may be logically consistent within the data it has, but contextually inappropriate.

Feedback Loops: In systems where AI output influences future input data, feedback loops can occur. For instance, in content recommendation algorithms, the system can get trapped in a loop of reinforcing its own biases, making it hard to detect errors in content filtering or recommendation.

Anomalous Data or Edge Cases: AI systems might struggle with data or situations that are significantly different from what they were trained on, known as edge cases. These anomalies can lead to errors that are hard to anticipate and detect because they occur outside the 'normal' operating parameters.

Errors in Complex Interactions: In systems where multiple AI models interact, errors can arise from the complex interplay between these systems. These errors can be particularly challenging to detect and diagnose because they may not arise from any single model but from the interaction between models.

Evolution of Language and Social Norms: AI systems trained on language data can become outdated as language and social norms evolve. Detecting when an AI system is no longer accurately interpreting or generating language according to contemporary standards can be difficult.

Misalignment with Human Values: AI might make decisions that are technically correct but misaligned with human ethics or values. These types of mistakes can be particularly hard to detect because they require nuanced understanding of human ethics, culture, and values.

To mitigate these issues, continuous monitoring, testing, diverse data sets, and incorporating explainability in AI systems are crucial. Additionally, involving human oversight and interdisciplinary expertise can help in detecting and correcting such hard-to-detect mistakes.

This page is intentionally blank

Chapter 6

Dark Data and Dazzle Data (Informative)

> *…we can be blind to missing data …that can lead us to conclusions and actions that are mistaken, dangerous, or even disastrous.*
> ***David Hand***

This page is intentionally blank

6.1 Introduction

Dark and dazzle data may be viewed as two sides of the same coin. In the former data is hidden, whilst in the latter the important information is hidden by an abundance of other data.

This page is intentionally blank

6.2 Dark Data

6.2.1 Introduction to Dark Data

This appendix outlines some of the safety implications arising from the issues identified by Prof. David Hand in his work on "dark data" presented in his book: "Dark Data: Why What You Don't Know Matters" [**?**] and website [**?**]. Dark data relates to data that is not available, but nevertheless is important, and indeed, in some cases, more important than the data that is available.

Unintentionally Donald Rumsfeld popularised the concept of dark data with his famous speech:

> *"…there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know…"*

Figure 6.1 neatly shows the dark varieties in grey:

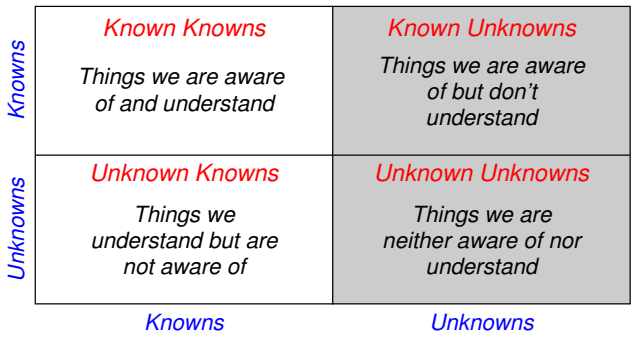


Figure 6.1: Varieties of dark data

6.2.2 Dark Data Example

A meeting was held prior to the fatal Challenger disaster in January 1986, to decide whether the flight should go ahead due to the prevailing cold weather and concerns over the performance of O-ring seals at low temperatures. The decision to launch was made on the basis of the following data points that show the temperature at which previous O-ring problems had occurred.

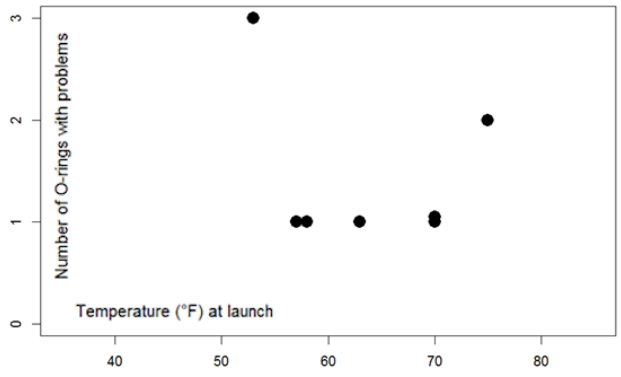


Figure 6.2: O-ring failures by temperature

On this basis, it was concluded that: "there is nothing irregular in the distribution of O-ring 'distress' over the spectrum of joint temperatures at launch between 53 degrees Fahrenheit and 75 degrees Fahrenheit." However, there was dark data in this dataset – the data points that show the temperatures for successful launches when there were no O-ring problems were left out. These are shown as the red points in Figure 6.3.

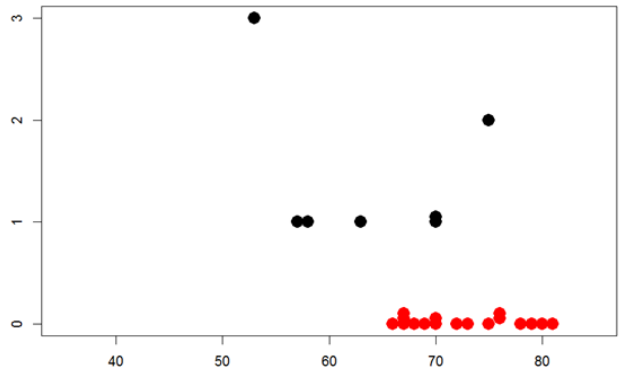


Figure 6.3: O-ring performance by temperature

When the Dark Data is included, there is an obvious correlation between temperature and O-ring failures and it is possible the decision to launch may have been different if this additional data had been presented.

6.2.3 Dark Data Varieties and Safety Examples

In his book, David introduces a taxonomy of 15 varieties of dark data. This section discuses each of these from a data safety perspective.

6.2.3.1Data We Know Are Missing: "Known Unknowns"

This case is very common in safety justifications where assurance information may be withheld for commercial reasons or does not exist, but we know, or are informed, that it isn't available. Common examples include:

- Assurance information for COTS components is unavailable
- Information about legacy systems was never produced or is now lost
- ML training data restricted to a particular context of use

If this is the case, it can be mitigated in several ways, including use of warnings, training, restrictions of use, etc. Evidence can also be substituted with other more indirect assurance e.g. established organisational track-record in the sector, or audit reports.

6.2.3.2Data We Don't Know Are Missing: "Unknown unknowns"

This is the most serious and far-reaching case. Occurrence is hopefully less common than case 6.2.3.1, but it is important to acknowledge that it does happen. Some examples are:

- The recent Covid-19 Track and Trace data loss (**??**), where the organisation handling the data was unaware that rows were missing from a spreadsheet for some time is illustrated in Figure 6.4
- Somebody knows a problem with a system but does not tell
- Machine learning training data missing edge / corner cases
- Key safety requirements missed
- Change of use never anticipated
- Data loss which may be discovered after some period (or indeed never)
- Test cases never thought of, so never created or executed

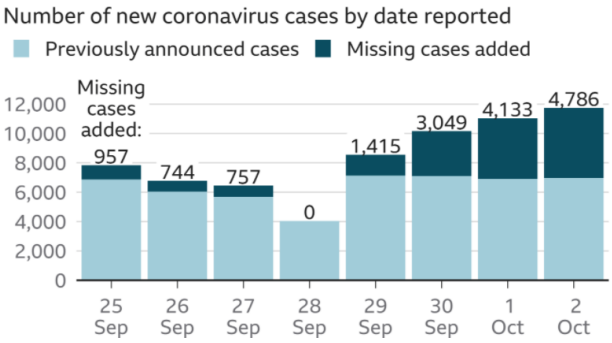


Figure 6.4: Covid-19 track-and-trace Data Loss

In many cases the loss may be discovered after some time, and it is incumbent on the organisation involved to analyse the impact of the missing data over the time period, including subsequent decisions and actions. Its effects should not be underestimated. This case can fundamentally change the safety picture and is probably in the highest safety risk category.

Data that was missing can subsequently be found. The following options show possible approaches to handling this "rediscovered data":

i) Apply the missing data

ii) Ignore the missing data

iii) Mention that the data was missing but not take it into account, and

iv) Perform an impact analysis on the missing data and then act according to the results

6.2.3.3Choosing Just Some Cases

This is where something or somebody has been selective. Examples might be:

- Selection of test runs that succeeded (ignoring failed runs and their diagnostics)
- Selective sampling from sensors, or where the sampling intervals are chosen badly
- Incorrect filtering of the data, leaving out more cases than intended

- Using data from completed forms, finished tasks or only including data which, intentionally or not, meets some criteria which excludes the important data. This can lead to "Survivorship Bias", [Wikipedia ref: https://en.wikipedia.org/wiki/Survivorship_bias] of which the following article has a good example from WW2 aircraft where examining the bullet holes on aircraft returning from missions could have led to the wrong conclusion: https://www.dgsiegel.net/talks/the-bullet-hole-misconception

Note that with complex or informal criteria the effects could be as case 6.2.3.2, i.e. you don't know what has been left out.

Mitigations include use of peer review, independent teams, and audits. It is important to ask questions and challenge the data selected to be sure it can be justified.

6.2.3.4Self-Selection

This is considered to be similar to case 6.2.3.3, but could be even more informal or ambiguous. Again mitigations include: use of peer review, independent teams, and audits.

6.2.3.5Missing What Matters

This was considered similar to case 6.2.3.2 in impact terms, and might well be considered the "Elephant in the Room". Examples might be:

- Measuring the wrong things, e.g. poor safety metrics / indicators
- Too much data to deal with or analyse, so some is ignored
- Too much filtering or processing, so losing information along the way
- Being too close to the data, i.e. the "wood for the trees". This is when the detail masks the overall issue with data, e.g. a slow trend or bias masked by peaks.

Mitigations include independence as 6.2.3.3 and 6.2.3.4 but also "taking a step back" to look at the bigger picture.

6.2.3.6Data Which Might Have Been

This is where it is impossible to obtain the relevant data as a consequence of how the system / scenario is constructed. This was considered an interesting and important case.

This could for example, be due to an inappropriate system architecture e.g. single data channel when a multiple channel approach should have been used, such as in the Boeing 737 MAX 8 accidents (see section H.2). It is therefore important that Mitigations are in place at design time, as they are often very difficult to retro-fit.

Mitigations include assessing the data collection opportunities gained or lost from the design or architectural approach – at design time.

6.2.3.7 Changes with Time

This is a common problem in a safety context. Data in safety systems often becomes obsolete or out of date and may still be mistakenly used. Some examples are:

- System configuration data not kept up to date as software or hardware changes
- Medical drug interaction databases
- Software patches requiring updates to configuration or system data, which is not always done

Mitigations include keeping critical data up to date and regularly refreshed, and knowing how old the data is in the first place.

6.2.3.8 Definitions of Data

This was considered a common case for systems with databases or those exchanging data with external systems, e.g.

- Taxonomies e.g. in machine learning categorisation schemes
- Data schemas in medical record systems
- Interface specifications and data definitions, which often evolve over time. These can render old data obsolete / subject to misinterpretation, and often needing migration / translation or re-interpretation.

Mitigations include keeping a list of known changes / incompatibilities and documenting the changes or fixes that have to be applied to make the data usable or consistent.

6.2.3.9 Summaries of Data

This is often seen in data about safety systems and projects:

- Safety metrics or indicators where data is aggregated to create a composite value
- Data fusion across multiple sensors Summaries can be misleading or cause "boundary reactions", e.g. Red-Amber boundary.

Mitigations include thinking about what missing data could cause (i.e. performing sensitivity analysis) – for instance, what decisions might be made erroneously due to certain data values being omitted, late or slightly perturbed.

6.2.3.10 Measurement Error and Uncertainty

Sensors can degrade and fail over time, especially in harsh environments such as automotive, marine or aviation:

- Sensors may feed faulty or biased data into systems

- Sampling techniques can cause some artefacts in themselves
- Interval polling interval can lead to misleading readings
- Data fusion can be impacted if multiple sensor values are combined

Hence sensors need to be regularly maintained and calibrated or replaced, and their interfaces need to be able to detect any problems and report faults.

6.2.3.11 Feedback and Gaming

This can often happen in safety justifications or test case production:

- Early production of a safety argument could lead to only those artefacts that support the claim being generated, e.g. requirements-based testing vs. stress testing
- Confirmation bias in safety justifications and can cause over-optimism

Mitigations include the use of dialectic argument approaches (where both positive and negative cases can be considered). It is also useful to get a second opinion, independent review or a 'fresh pair of eyes' to check the justification.

6.2.3.12 Information Asymmetry

This is common where there are multiple stores or sources of the same data:

- Multiple / backup databases where they are not kept in sync
- Image recognition systems that learn, but fail to share data
- Issues of divergence of data across multiple sources.

Comparisons, reviews and data audits may help in these cases. In general the issue of divergence across multiple data sources needs to be recognised and addressed in the most appropriate technical, and ideally automated, way.

6.2.3.13 Intentionally Darkened Data

This can and does happen, e.g.

- Defence, security and government sectors where data is purposefully hidden or destroyed
- Records intentionally deleted after an accident to cover up what really happened

This can be mitigated using appropriate technical measures (e.g. blockchain, digital signatures, off-site backups), and also robust procedures with oversight.

6.2.3.14 Fabricated and Synthetic Data

Fabrication (i.e. making up the data) is surprisingly common, e.g. in medical, policing and maritime sectors. Sometimes it is created with the best of intentions as it was not produced or collected at the right time and is now required (e.g. for an audit), but sometimes it is created to mask a problem. Synthetic data is often used where there are difficulties in producing enough real data with the right characteristics:

- Data is retrospectively entered / patched to make a "clean" record
- Synthetic autonomous vehicle training databases can have issues with artificial data if not realistic

Possible mitigations include comparisons with current real-world data and checking with historical data.

6.2.3.15 Extrapolating Beyond Your Data

Systems, especially machine learning systems (see Appendix J) have to cope with values outside of their training data, but the outcomes may be unexpected:

- Bizarre results from recognition / detection systems may result

Mitigations include use of machine learning training data containing edge / corner cases, boundary cases, and testing well beyond normal or acceptable ranges to give insight into the system degradation behaviour.

6.2.4 Summary of Dark Data

Dark data is a very useful way of thinking about problems and solutions. It is important to always think of the bigger picture, considering what information is not known:

- What might have been left out, intentionally or otherwise?
- What might be missing due to the way things are done (or the way those things are being executed)?
- Could missing data items cause any safety problems?
- What is outside of the defined operational domain?
- Can you present your results / outputs in a way that shows the dark data?

6.2.5 Further Reading

A number of resources are available for further information on how dark data relates to data safety:

- The October 2020 edition of the SCSC Newsletter [**?**] contains articles both by David Hand and on dark data safety.
- A presentation on dark data given by David Hand to the DSIWG [**?**];
- A presentation by Mike Parsons given at the University of York [**?**].

This page is intentionally blank

6.3 Dazzle Data

6.3.1 Introduction to Dazzle Data

This appendix outlines some of the safety implications arising from the issues identified by "Dazzle data" in contrast to "dark data" presented in David Hand's book: "Dark Data: Why What You Don't Know Matters" [**?**] and website [**?**]. Dark Data relates to data that is not available but nevertheless is important.

"Dazzle data" refers to spurious, superfluous or unexpected data that masks and confuses the picture and can reduce your ability to see details, and indeed the whole picture – a bit like noise or camouflage.

> It is data you don't want, didn't ask for, arrives more frequently than expected, is redundant, or data that arrives at an unexpected or inconvenient time or in the wrong format or sequence, so requiring extra resources to deal with…

At first sight it might be thought this is just an annoyance rather than a problem, however Dazzle data can mask, distort or prevent use of the real or desired data. It can hide the true picture as in "wood for the trees" and in the worst case create real problems.

Some examples of such false positive alerts that cause real alerts to be ignored are:

- The classic case of "Crying Wolf".
- Denial of service attack, when servers can be blocked with useless requests that cannot be distinguished from the real requests.

Data which is truly spurious and believable (e.g. generated by extensive electrical noise, or fabricated to mask a fraud) may be accepted as valid data, leading to potentially hazardous results.

Varieties of dazzle data may be identified in the same grid manner as Dark Data, as shown in Figure 6.5, where the yellow areas indicate potential dazzle data.

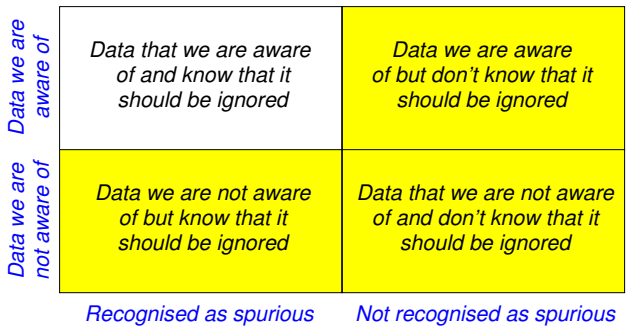


Figure 6.5: Dazzle data varieties

6.3.2 Dazzle Data Examples

6.3.2.1Data We Are Aware Of – Recognised as Spurious

Repeated false positive alerts (e.g. from faulty sensors or alarms). These tend to be annoying irritants rather than safety issues, however they can lead to alarms being permanently switched off, masking true problems. An example from everyday life might be removing the battery from a badly place smoke alarm in a kitchen which repeatedly sounds when the toaster is used.

6.3.2.2Data We Are Aware Of – Not Recognised as Spurious

Electromagnetic interference (EMI) affecting transmitted data is a good example of this case; we know it exists as a possibility but don't know how it might affect data in a system or a communications path, and this could be in different ways at different times or days.

6.3.2.3Data We Are Not Aware Of – Recognised as Spurious

In many cases we do not know what changes or corruptions may affect the data, but we have strong correction mechanisms to cope with the problems regardless. A good example is data stores on a spacecraft which have redundancy and error correcting codes built-in and so can deal with the majority of corruptions caused by cosmic rays.

6.3.2.4Data We Are Not Aware Of – Not Recognised as Spurious

An example of this is where data is mistakenly added to a machine learning training dataset resulting in bizarre and possibly hazardous behaviours of the system using it (e.g. an autonomous road vehicle). It is paramount that data used for such image recognition systems filters and excludes potentially dangerous extraneous data arising from unanticipated situations

6.3.3 Dazzle Data Varieties and Safety Examples

In his book, David Hand introduces a taxonomy of fifteen varieties of Dark Data. This section discuses each form of Dazzle Data and considers it from a systems and safety assurance perspective. Some of the varieties are the same as Dark Data, but some are very different.

6.3.3.1Data We Know are Superfluous or Unneeded: "Known Extras"

This case is very common in systems where an interface may be flooded with unsolicited messages. It is also common in safety justifications where assurance information may be padded out with extra, irrelevant information that can mask the underlying safety argument (which may or may not be intentional). It can be a serious problem in large, complex safety documents or detailed safety analyses. Luckily, in this case, the data is recognised as superfluous and can be ignored. Common examples include:

- Mailboxes filled with spam mail, preventing important messages being read
- Repeated stall warnings when the aircraft is under control

- Assurance information in a safety case is included for components which are not part of the solution in use at that site or situation
- Assurance information about old or legacy components is included but these are now not in service
- Machine learning training data contains many cases which are duplicates or very closely related, so adding nothing to the learned behaviour but making the dataset hard to deal with

This case can be mitigated in several ways, including careful filtering, review, use of concise notations (such as Goal Structuring Notation) to extract the essence, and periodic audits and cleans.

6.3.3.2Data We Don't Know Are Unneeded or Spurious: "Unknown Extras"

This is the most serious and far-reaching case, where the additional data is not recognised as unneeded and may be processed, analysed or be left in place when it should be removed. Some examples are:

- Legacy code in software where nobody understands its function, or how it relates to the current code and so is reluctant to remove it. The Ariane 4 Inertial Reference function left in place in the Ariane 501 launch disaster might be such a case. Note that code whose function is known would be under the first case, discussed in 6.3.3.1
- Parts of the safety case or safety argument are supplied separately (e.g. by a subcontractor or 3rd party) and they are obscure. In this case it will not be known which parts are relevant to the particular situation. Some information may be irrelevant or worse, misleading.
- Machine learning training data containing too many outliers, which are not recognised as such

In many cases the extra data may be discovered after some time, and it is incumbent on the organisation involved to analyse the impact of the complete set of data which may have been used over the time period, including subsequent decisions and actions. The effect of spurious data should not be underestimated, as it may have masked, prevented or distorted the intended use of the nominal data for some time. This case can fundamentally change the safety picture and is probably in the highest safety risk category.

Data found to be erroneous (i.e. it should not ever have been in the system or in the safety argument) is a problem. The following options show possible approaches to handling this data:

i) Delete the erroneous data

ii) Accept the erroneous data, i.e. continue using it

iii) Report that the data was erroneous but take no further action

iv) Establish under what circumstances and situations the erroneous data would have had an effect

v) Perform a full impact analysis on the effects of the erroneous data and then act according to the results

6.3.3.3Data Obscuration: Missing What Matters

This is where the meaning of the overall dataset becomes obscured due to the extra unnecessary data. Examples might be:

- Measuring the wrong things due to excessive noise or too much data to deal with
- Processing involving sampling only picks bad data elements
- Being too close to the data, i.e. the "wood for the trees". This is when the extra data masks the overall issue with the data, e.g. a slow trend or bias hidden by large local variations.

Mitigations include review, statistical checks and "taking a step back" to look at the bigger picture.

6.3.3.4Data Masking in Specific Cases

This is where the extra data specifically masks, obscures or hides particular data elements (but not all). A particularly nasty case of this is where filters are put in place to remove unwanted data values, but those filters actually remove less (or more) than they should (i.e. don't remove all unwanted cases or remove valid values as well). Some examples might be:

- The number of successful test runs vastly outweighs failed runs and so the failures are not investigated
- Incorrect filtering of the data, leaving in some cases that should have been excluded

Mitigations include use of review and completeness checks. Note that over-aggressive filtering would create cases of Dark Data.

6.3.3.5Masquerade or Fraudulent Data

This is where data has been constructed to fool the system consuming it, hiding, overlaying or replacing the correct data. Often this will be malicious and should be filtered or rejected by the target system, but of course may not be.

- Intentional fraud
- Some security attacks

Mitigations include audit, blockchain, monitoring, intelligent profiling of data and detection of changes.

6.3.3.6Incorrect Definitions of Data

This is considered a common case for systems exchanging data with external systems, where they may duplicate, translate or incorrectly send multiple messages. They may be time-separated, for instance if communications are lost and then regained, leading to confusion.

- Data sent in the wrong rate or units (e.g. every millisecond rather than second)
- Data given for the wrong range or duration (e.g. for a month rather than a day
- Data sent for the wrong domain or scale (e.g. national values rather than regional)
- Text messages re-sent to a mobile phone when coverage restored

Mitigations include making sure interface specifications are clear and ambiguous, rejecting incorrect size or scale datasets, keeping a list of known changes / incompatibilities and documenting the changes or fixes that have to be applied to make the data usable or consistent.

6.3.3.7 Summaries of Data

Spurious data can affect summaries in a significant way if composite values are calculated, e.g.

- Calculations of averages or other statistics affected by duplicates
- Spurious outliers can lead to over fitting when analysing data trends

Mitigations include audit and assessing what extra data could cause (i.e. performing sensitivity analysis) and establishing what decisions might be made erroneously due to certain data values being present.

6.3.3.8 Information Asymmetry

This is common where there are multiple stores or sources of the same data. If one has erroneous duplicate values (i.e. many null values used for padding) then comparisons between them may fail.

- Multiple / backup databases where they are not kept in sync
- Issues of divergence of data across multiple sources.

Comparisons, reviews and data audits may help in these cases. In general the issue of divergence across multiple data sources needs to be recognised and addressed in an automated way.

6.3.3.9 Intentionally Dazzled Data

This might happen in a secure context where key data is effectively "camouflaged" by hiding within large data items (e.g. images) or the use of large amounts of apparently routine data to mask the secure data.

- Defence, security and government sectors where data is purposefully hidden
- An organisation may provide copious amounts of safety case collateral to hide a weak safety case

6.3.3.10 Extrapolating Beyond Your Data

Machine learning systems have to cope with values outside of their training data, but the outcomes may be unexpected if the training data contains many spurious values:

- Bizarre results from recognition / detection systems due to out of range values
- Image components mislabelled leading to strange results

Mitigations include checking of the values that are used in training, analysis of outliers and repeats, deletion of duplicates and use of machine learning validation data containing edge / corner cases and boundary cases.

6.3.3.1 A Note on Sensors

Sensors can degrade and fail over time, especially in harsh environments such as automotive, marine or aviation. In such cases sensors may feed erroneous or additional data into systems (e.g. if an out of normal range situation is detected), leading to misleading processing or false alarms. The Boeing 737 MAX 8 accidents might fall into this category as the angle-of-attack sensor erroneously generated high values, and certainly the systems / processing which passed on the values created spurious data. Some examples are:

- Faulty sensors
- Multiple sensors that do not have their values properly combined
- Sampling techniques which generate additional data
- Interval polling interval can lead to misleading readings
- Data fusion can be impacted if multiple sensor values are combined

Mitigations include independent monitoring of sensors and regular maintenance, calibration or replacement of hardware. Their interfaces need to be able to detect any problems and report faults. In particularly critical applications, it may be necessary to minimise the potential for common mode failures by using disparate technologies, or at least sensors from different manufacturers.

An example may be found in https://en.wikipedia.org/wiki/Qantas_Flight_72 where:

> ...the CPU[1] of the ADIRU[2] corrupted the AoA[3] data. The exact nature of the corruption was that the ADIRU CPU erroneously re-labelled the altitude data word so that the binary data that represented 37,012 (the altitude at the time of the incident) would represent an angle of attack of 50.625 degrees. The FCPC[4] then processed the erroneously high angle of attack (AoA) data, triggering the high-AoA protection mode, which sent a command to the electrical flight control system (EFCS) to pitch the nose down.
>
> The FCPC algorithm was very effective, but it could not correctly manage a scenario where there were multiple spikes in either AoA 1 or AoA 2 that were 1.2 seconds apart – i.e., if the 1.2-second period of use of the memorised value happened to end while another spike was happening.

6.3.4 Summary of Dazzle Data

Dazzle data is a very useful way of thinking about data problems and solutions. It is important to always think of the bigger picture, considering what information is superfluous and may be distorting or masking the real picture:

- What might be hidden intentionally or otherwise due to the extra values?
- Could the extra data items cause any safety problems, e.g. change a numerical calculation?
- How would a system cope if it received data it was not expecting?

1 Central processing unit
2 Air data inertial reference unit
3 Angle of attack
4 Flight control primary computer

- How do you characterise the nature of unexpected data so as to ensure it can be handled if it does occur?
- Can you present your results / outputs in a way that shows the extra data?

This page is intentionally blank

6.4 Links between dark data / Dazzle Data and

Fundamentally, dark data is data that is not available, but nevertheless is important, and indeed, in some cases, more important than the data that is available.

Similarly, Dazzle data refers to spurious, superfluous or unexpected data that masks and confuses the picture and can reduce your ability to see details, and indeed the whole picture – a bit like noise or camouflage.

The opposite attributes of dark data and dazzle data might be termed Brightness and Clarity respectively. These are potential new properties or meta-properties of data.

Table 6.1 illustrates the properties which can be affected by dark data and dazzle data issues.

Table 6.1: Data properties affected by dark data and / or dazzle data issues

Data property	Explanation	Dark Data – missing data	Dazzle data – obscuring data	Notes
Integrity (I)	The data is correct, true and unaltered	✓	✓	E.g., Corruptions to a database could change the original values so the original data is lost, or hide or mask values such as by inserting End-of-File markers
Completeness (C)	The data has nothing missing or lost	✓	-	If data is lost then it is usually unknown and hence dark. However use of parity, cyclic redundancy checks (CRCs) and digital signatures may be able to detect the loss and, in some cases, fill in the missing data.
Consistency (N)	The data adheres to a common world view (e.g., units)	-	-	
Continuity (Y)	The data is continuous and regular without gaps or breaks	✓	-	Gaps or breaks in a data stream would be dark – the missing values are unknown. Methods may be available to mitigate if detected, e.g., interpolation.
Format (O)	The data is represented in a way which is readable by those that need to use it	✓	✓	If data is not in the correct format then it may not be usable so becomes dark. If lots of data is in the wrong format it may obscure other (good) data, and hence can dazzle.
Accuracy (A)	The data has sufficient detail for its intended use	✓	-	If the data has lost detail (e.g., due to sensor sampling or representation) then this is a dark case.

Continued on next page

Table 6.1: Data properties affected by dark data and / or dazzle data issues (continued)

Data property	Explanation	Dark Data – missing data	Dazzle data – obscuring data	Notes
Resolution (R)	The smallest difference between two adjacent values that can be represented in a data storage, display or transfer system	✓	✓	If the system cannot distinguish between different data values that actually are different, then data could become hidden or lost and hence dark. If the data is masked then it may also be a case of dazzle data.
Traceability (T)	The data can be linked back to its source or derivation	✓	-	If traceability is lost then the data effectively becomes darker as its origins and derivation are lost, and may not be able to be used.
Timeliness (M)	The data is as up to date as required	✓	✓	If data is not, or thought to be not, up to date then it may be rejected and hence become dark data. But clocks can become out of sync across different systems leading to erroneous rejection so the timestamps may then be thought of as dazzling the system.
Verifiability (V)	The data can be checked and its properties demonstrated to be correct	✓	-	If data can't be checked then it may be rejected (and become dark), even if it is valid.
Availability (L)	The data is accessible and usable when an authorised entity demands access	✓	-	If data is not available as required then it is dark (whether it exists or not).
Fidelity / representation (F)	How well the data maps to the real-world entity it is trying to model	-	-	
Priority (P)	The data is presented / transmitted / made available in the order required	✓	✓	If data is not prioritised to be available when needed it is effectively dark. This could be because other data is taking precedence so could be a dazzle case.

Continued on next page

Table 6.1: Data properties affected by dark data and / or dazzle data issues (continued)

Data property	Explanation	Dark Data – missing data	Dazzle data – obscuring data	Notes
sequencing (Q)	The data is preserved in the order required	✓	✓	If data is not sequenced as needed it may be rejected or ignored and then is effectively dark. Other data which is out of order or sequence may cause the whole message, set or file to be rejected – a dazzle case.
Intended destination / usage (U)	The data is only sent to those that should have access to it	✓	-	If data is not sent to the correct destination it may be a dark data situation as it is effectively missing.
Accessibility (B)	The data is visible only to those that should see it	-	✓	If data is more widely visible or distributed than intended it could cause a dazzle situation due to e.g. bandwidth issues.
Suppression (S)	The data is intended never to be used again	-	✓	If data re-emerges after it should have been deleted it could cause all sorts of dazzle-related problems.
History (H)	The data has an audit trail of changes	✓	-	If the audit trail is lost then the data is darkened as its origins and derivation could be important.
Lifetime (E)	When does the safety-related data expire	✓	✓	If the lifetime erroneously ends too early then the data is lost – a dark data problem. Conversely if the lifetime is too long it is potentially a dazzle situation.
Disposability / deletability (D)	The data can be permanently removed when required	-	✓	If the data cannot be removed it is potentially a dazzle situation, as it may obscure the real or current data.
Goldilocks (G)	The data is just the right size — not too much and not too little	✓	✓	Too little data may be a dark data case and too much data to deal with is a dazzle case.
Analysability (Z)	The data is of a suitable size, type and representation (including any metadata) to enable it be usefully analysed	✓	✓	If the data can't be analysed it may not be able to be used causing a dark data issue. If there is too much of it or it is obscured (e.g., by noise on a comms line) then it is a dazzle case.

Continued on next page

Table 6.1: Data properties affected by dark data and / or dazzle data issues (continued)

Data property	Explanation	Dark Data – missing data	Dazzle data – obscuring data	Notes
Explainability (X)	The data can be meaningfully explained to those who need to understand it by a suitable mechanism	✓	✓	If the data can't be explained it may not be able to be used causing a dark data issue. If there is too much of it or suitable management and analysis tools are not available then it is potentially both cases.

Chapter 7

Data Cygnology (Informative)

> *Here be dragons*
> ***Anon***

This page is intentionally blank

7.1 Introduction

The term "Black Swan" is used to describe an unpredictable event that is beyond what is normally expected of a situation and has potentially severe consequences. Black Swan events were originally discussed in a financial context by Nassim Taleb in his 2004 book *"Fooled By Randomness"* [**?**] and further expanded to cover other topics in *"The Black Swan: The Impact of the Highly Improbable"* [**?**]. The term quickly became accepted, and Black Swan events are characterized by their extreme rarity, severe impact, and the widespread insistence they were obvious in hindsight. Taleb gave examples such as the rise of the Internet, the personal computer and financial crashes as "Black Swan" events, but the term can equally be applied to unexpected safety-related events such as the Boeing 737 MAX crashes of 2018 / 2019 [**?**].

While Black Swans are events that are only obvious in hindsight, other types of event may be extremely rare and have severe impact but are predictable beforehand, for example pandemics. Another example comes from the sun: a Coronal Mass Ejection (CME) of such severity can be produced that it causes widespread life-threatening electromagnetic interference on Earth. This type of event can be predicted to occur to some extent, and has happened before, for example the "Carrington Event" of 1859 [**?**]. These rare events which have rapid escalation, but are somewhat predictable are termed "Dragon King" events.

There are several more event types noted in the literature. The practice of studying these types of high-impact but rare variants has been termed "Cygnology", a term that appears in work by Ale, Hartford and Slater in their "Dragons, black swans and decisions" paper of 2020 [**?**].

While Cygnology can be applied to safety in general, it can also be applied to data safety and recent research has identified four important categorisations of data-related risks:

- Black Swan Data
- Dragon King Data
- Perfect Storm Data
- Pudding Lane Data

These types of data are discussed in more detail in the following sections along with recommended strategies for managing their associated risks.

This page is intentionally blank

7.2 Black Swan Data

Definition: Data that is totally unexpected by those receiving it (i.e. 'out of the blue') and has huge (detrimental) impact, but in retrospect should have been anticipated.

In 2013, a C-130 Hercules aircraft landed at Bar Yehuda airport near the Dead Sea, a saltwater lake sitting astride Israel and Jordan. The Dead Sea is the lowest place on Earth and the airfield lies -1,210 feet below sea level.

The aircraft's navigation system became unresponsive and the constellation of Global Positioning System (GPS) satellites above, mysteriously winked out of existence. As it turned out, the plane's navigation electronics were not designed to operate at altitudes less than 400 feet below mean sea level. In a sense, the plane thought it was underwater.

This is an example of what is termed a "Black Swan Data" event; the stream of altitude data going to the navigation systems surprisingly turned significantly negative and could have had a catastrophic result; a situation not considered possible. Yet, with hindsight, and knowledge of earth's geography, we can rationalise this and question why the assumption of positive altitude data was ever thought to be true in the first place.

7.2.1 Managing Black Swan Data Risks

The following strategies are recommended for managing Black Swan Data Risks:

- Be aware of changes of usage, environmental conditions or failures that may produce data that is not expected. The Ariane 501 launch disaster may be thought of in these terms, as the horizontal bias values that caused the exception leading to the complete loss of mission were not anticipated;
- Continually test the established thinking; challenge the oft quoted "that's just how we do things around here";
- What assumptions have been made about the data? Are these really valid and do they continue to be valid in a changing and dynamic world?
- What happens if different inputs (e.g. new sensors or information flows) are added to your system?
- Think outside the box; it is the most wild and implausible (compared to current perceptions) ideas that will trip a Black Swan Data event – what could possibly be 'bad data' and how might it be handled?

This page is intentionally blank

7.3 Dragon King Data

Definition: Data whose effect might have been foreseen, but leads to an unexpected and explosive escalation with major impact ('things get rapidly out of control').

In risk management a Dragon King is defined as an event that is both extremely large in size or impact (a "king") and born of unique origins (a "dragon") relative to other events in the system.

Dragon King events are generated by, or correspond to, mechanisms such as positive feedback, tipping points, bifurcations, and phase transitions that tend to occur in non-linear and complex systems, and serve to amplify events to extreme levels. In the data domain, "Dragon King Data" represents incorrect or missing data whose consequences unexpectedly escalate into an unpredicted and often catastrophic system event.

An example might be the loss of data relating to Covid-19 test results in the UK due to a row limit in Microsoft Excel. This led to Public Health England losing 15,841 positive test results, which in turn, meant that 50,000 potentially infectious people were missed by contact tracers and not told to self-isolate. Because of the exponential way that viruses spread, this may have led to many more infections and deaths.

Another, older, example is the infamous comment given by the weather forecaster Michael Fish about the Great Storm of 1987:

> "Earlier on today, apparently, a woman rang the BBC and said she heard there was a hurricane on the way. Well, if you're watching, don't worry, there isn't!"

These few words may well have led to precautions not being taken across the whole of Southern England, affecting millions of people. In fact, the storm was the worst to hit South East England for three centuries, causing record damage and killing 19 people.

7.3.1 Managing Dragon King Data Risks

The following strategies are recommended for managing Dragon King Data Risks:

- Identify all the receiving systems or users of the data: what are the consequences of errors or losses in the data to those receiving it?
- Analyse the effects of data impact: are any effects subject to non-linear effects? Try to establish the consequence chains and knock-on effects;
- Do any of the data potentially impact systems that have large fan-out or spread?
- Create models (or even Digital Twins) of the system in a way that different combinations of data can be tested and simulated to reveal extreme non-linear behaviours.

This page is intentionally blank

7.4 Perfect Storm Data

Definition: Combinations, sets or occurrences of data (or absences of data) that would never have been thought possible to occur together, and when they do have a large and undesirable impact.

A Perfect Storm in the risk world is when everything bad happens at the same time, and it was not anticipated that this could happen. These risks get their name from "The Perfect Storm", both a 2000 American biographical disaster drama film [**?**] and a 1997 non-fiction book by Sebastian Junger [**?**]. In the data world, a Perfect Storm is when the data that the system relies upon are all wrong (or missing) at the same time, and can be layers deep – much akin to Reason's Swiss cheese model. A classic example might be to try and restore deleted data from a backup system, only to find that the backups had not been working properly due to a configuration problem. Perfect Storm Data is therefore when multiple data faults or losses conspire to create a situation that has disproportionately catastrophic outcomes.

Another example is from the world of security; on 10th Dec 2021, a new critical zero-day vulnerability was detected that affected Apache Log4j 2 Java library. It adversely impacted the digital domain and security systems worldwide. The vulnerability, when exploited, permitted remote code execution on the vulnerable server with system-level privileges. The exploit was a combination of both the Java code that contains different logging functions and the settings of a configuration file.

Another example is the China Airlines A333 at Taipei on 14th June 2020 when all primary computers, reversers and auto brakes failed on touchdown. During landing, flight controls reconfigured from 'normal law' to 'direct law' after all three flight control primary computers (FCPCs) became inoperative. While all aircraft primary control surfaces were still controllable, the deceleration devices including ground spoilers, thrust reversers, and autobrake were lost; the deceleration of the aircraft had to be performed by manual braking by the pilots.

7.4.1 Managing Perfect Storm Data Risks

The following strategies are recommended for managing Perfect Storm Data Risks:

- Analyse combinations of data that are important: how might one dataset being wrong influence another?
- Look at dependencies in the data: are seemingly different datasets all derived from a common source?
- Examine as many levels of the system to establish commonalities and dependencies;
- Even if no commonalities are found, consider the unlikely cases of independent data being wrong together, and follow through their consequences.

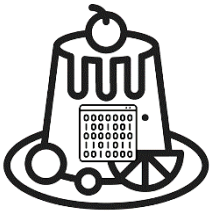
This page is intentionally blank

7.5 Pudding Lane Data

Definition: Data that is unknowingly and unexpectedly critical to the whole operation and if missing or incorrect in some way, has a dramatic and negative impact. After the event it may be obvious that it was critical.

> "Lord! what sad sight it was by moone-light to see, the whole City almost on fire, that you might see it plain at Woolwich, as if you were by it".

These words from Samuel Pepys diary in 1666, give a personal account of the Fire of London that destroyed 13,200 houses, 87 parish churches, The Royal Exchange, Guildhall and St. Paul's Cathedral, all from a single fire thought to have started in a humble baker's shop in Pudding Lane.



In the same way, unexpectedly critical data can also act as a catalyst to catastrophic outcomes if the dependencies are not identified and adequately mitigated. Often this data is part of an existing system and may be long forgotten (e.g. a system configuration parameter). It may also be a hard-wired parameter in legacy source code that code maintainers change because they do not understand its original purpose.

In 2015, an Airbus A400M Atlas cargo plane on a test flight crashed at La Rinconada, Spain, less than 5 kilometres from Seville Airport, killing 4 of the 6 crew. Several reports suggested that as many as three of the aircraft's four engines failed during the A400M's departure from Seville.

The suspected cause of the failure was that the torque calibration parameter data had been accidentally wiped on three engines as the engine software was being installed, which would prevent the engine control software from operating properly. In this case, the seemingly innocuous calibration data had a critical role to play in the safety of the entire platform and so the term "Pudding Lane Data" is coined to encompass this form of data.

7.5.1 Managing Pudding Lane Data Risks

The following strategies are recommended for managing Pudding Lane Data Risks:

- Understand and know all the data. The high-profile data streams will of course be first and foremost, but is there an inventory of all the data that the system depends upon and are the implications if there is loss of their data properties understood?
- Configuration and tailoring data may seem static and uninteresting but what are the dependencies on this data? If loss of properties of this data can have catastrophic outcomes, then what processes and mitigations will be put in place to assure those properties are maintained?
- Consider the reporting and monitoring data: if this is incorrect will poor decisions be taken?

This page is intentionally blank

7.6 Conclusions

Looking at Data Risks through Cygnology can help to tease out some unlikely but severe failures of systems involving data. It is possible that sometimes these risks are already considered in system HAZOP sessions but discounted as too unlikely to happen. The Data Cygnology approach is recommended as a useful and visual 'Tool in the Box' when trying to identify extreme data-related risks of systems. Due to their serious impact arguably they should still be included, assessed and managed alongside more 'normal' risks.

This page is intentionally blank

7.7 Summary Table

Table 7.1: Summary of cygnology types

Type	Brief Explanation	Icon
Black Swan Data	Data that is totally unexpected by those receiving it (i.e. 'out of the blue') and has huge (detrimental) impact, but in retrospect should have been anticipated.	
Dragon King Data	Data whose effect might have been foreseen, but leads to an unexpected and explosive escalation with major impact ('things get rapidly out of control').	
Perfect Storm Data	Combinations, sets or occurrences of data (or absences of data) that would never have been thought possible to occur together, and when they do have a large and undesirable impact.	
Pudding Lane Data	Data that is unknowingly and unexpectedly critical to the whole operation and if missing or incorrect in some way, has a dramatic and negative impact. After the event it may be obvious that it was critical.	

This page is intentionally blank

Chapter 8

The Data Safety Tool: RADISH (Discursive)

> *Technology is nothing. What's important is that you have a faith in people, that they're basically good and smart, and if you give them tools, they'll do wonderful things with them.*
> ***Steve Jobs***

This page is intentionally blank

8.1 An introduction to the tool

RADISH (The Risk Assessor for Data Integrity and Safety Hazards) is a software tool being developed by Mission Critical Applications Limited (mca-ltd.com), to assist a data safety practitioner developing a data safety case using the guidance, by:

- recording the decisions that are made;
- automating parts of the data safety assessment process; and
- helping the practitioner to choose between the risk mitigations that are recommended by the guidance, given the nature of each risk.

For more information, and to access to the RADISH tool, visit data-safety.tech/tooling.

As illustrated in Figure 8.1, the RADISH tool is a central repository of information about the data safety case for a project. RADISH guides data safety engineers to identify the data artefacts in the system, and the safety properties that are important for each artefact. It then calculates the highly recommended and recommended mitigation techniques following the guidance, allowing the engineer to chose those which should be in the system requirements those which should not, and record the justifications for those choices.

RADISH can generate reports showing the risks that have mitigations, and those where more work is needed, giving project management a view of the state of the data safety process. RADISH can also generate a report that can be included in a data safety case to support the system safety argument, and this can continue to evolve throughout the project lifecycle of development, deployment and maintenance.

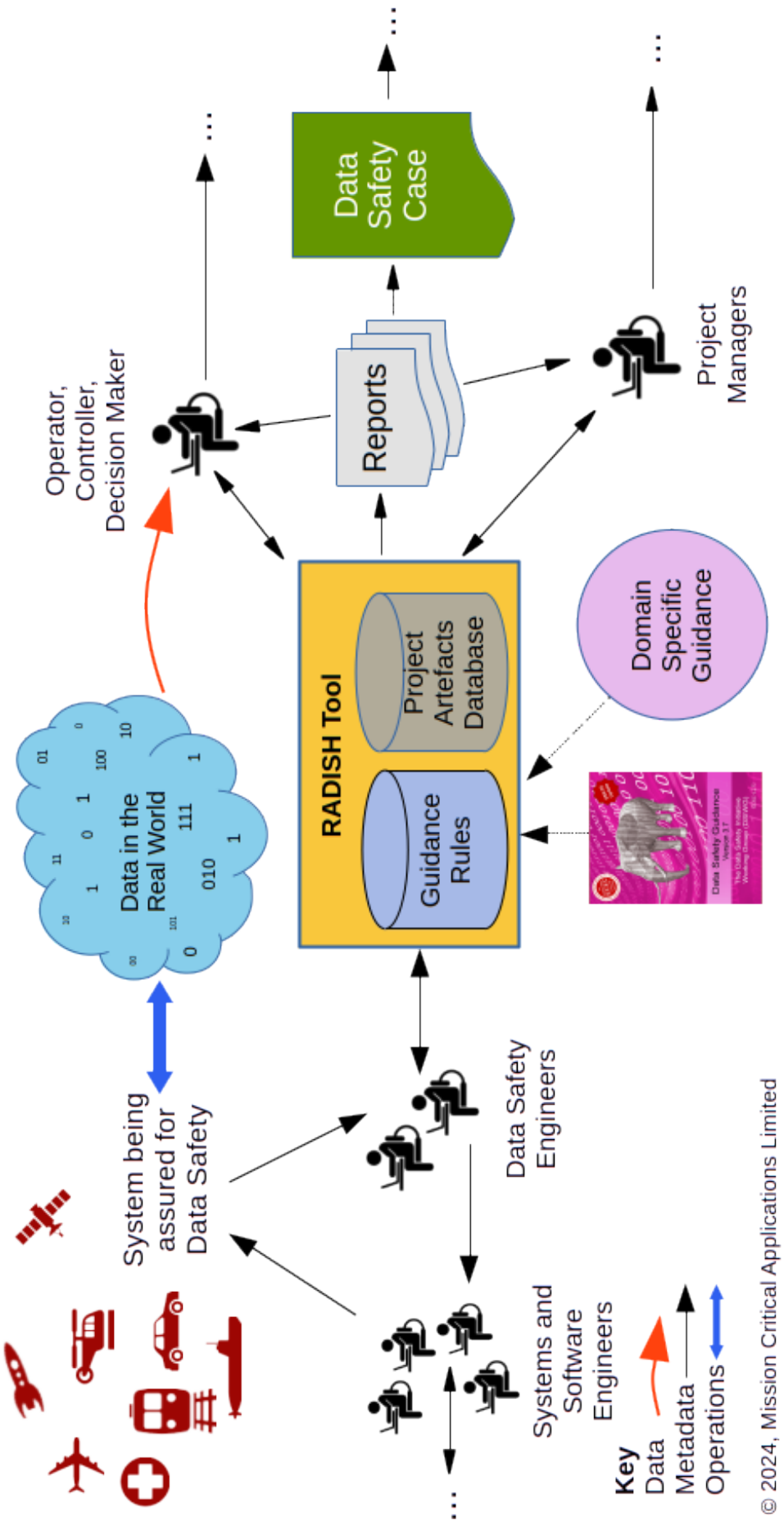


Figure 8.1: RADISH within the project environment

8.2 A caveat

The tool described in this appendix is usable, but is not yet at production standard. The developers are working towards a release in 2025, and are looking for programmes to use as test vehicles. To have a programme considered as a possible test vehicle for the tool, please contact the developers at Mission Critical Applications Limited (radish@mca-ltd.com).

This page is intentionally blank

Chapter 9

Covid-19 (Informative)

> *The public health community wants a safe and effective [COVID-19] vaccine as much as anybody could want it. But the data have to be clear and compelling.*
> ***Michael Osterholm***

This page is intentionally blank

9.1 Covid-19 and Data

The Covid 19 crisis has highlighted a number of areas where better data and data management could have improved outcomes and hence reduced the death toll. Such pandemic-related data is therefore very much safety-related data.

Some issues related to Covid-19 data are:

1. The lack of consistency and standardisation in handling the data, e.g:
    a. In predictive models where many assumptions may be wrong, or the algorithms inappropriate
    b. Presentation of statistics in a selective or misleading way
    c. Methods of data collection which may be selective or incomplete. For instance reporting on the number of positive test cases is always misleading, as many people may be asymptomatic with the virus so never get tested.
    d. Calculations and filtering
    e. Allowances for delays in collection or processing
    f. Intentional and unintentional bias
    g. Use of averaging (e.g. moving averages) and smoothing of plots hiding sudden increases
    h. Loss of data (e.g. in the recent UK Test and Trace system due to old versions of Excel)

    All of which have prevented any meaningful comparisons internationally, even between countries in Western Europe. They also lead to public confusion; this in turn leads to mistrust and a refusal to abide by guidance and regulations.
2. The poor data within the Test and Trace systems is a major factor in the failure of these systems. If data is not accurate, timely or complete then contacts cannot be traced in time, and the effort put into the activity is wasted. These systems in the UK need huge improvement as there is currently low contact performance and hence very poor outcomes. Background reading on this may be found at this link: https://www.bbc.co.uk/news/health-55008133

This page is intentionally blank

9.2 Systems Involved with Covid-19 Data

Many systems have been created or re-deployed to help manage the pandemic. These systems consume and produce vast amounts of data, some of which is critical and could affect safety of individuals or the general population. Some identified systems are shown in Table 9.1. For each of these, it is worth thinking through some basic data failure modes, e.g. data is lost, late, incorrect or incomplete. For instance if we are running an infection / spread model and we feed it with stale data, then its predictions will clearly be inaccurate.

Table 9.1: Systems involving data used to manage the pandemic

Analysis of air flow and particles	Satellite imagery (Wuhan)	Video conferencing	Risk assessment systems
Infection / spread models (inc new variants)	Itinerary systems	Remote consultation systems	Computational bioinformatics tools
Infra-red / thermal cameras	Infection testing systems	Ventilators / other patient management devices	Appointment systems
Track and Trace apps	Antibody testing systems	Personal risk profiling apps / systems	Border Control / Quarantine systems
Track and Trace back office systems	Drug trials systems / data	Allocation / reservation / booking systems	Risk profiling / prioritising for vaccination
Track and Trace service	Ventilation models & UV sterilisers	Models of built environments	Digital Twins (systems and biological: lungs, etc.)
Supply chain systems	Behavioural models	Safety analyses (STAMP/STPA), etc.	(Automatic) cleaning systems
Virus aerosols modelling	Analysis of delays in system of reporting / actions	Modelling / public perception of the disease	Virus shedding models
Vaccination booking / tracking / monitoring	Sanitiser systems	Lockdown easing models	Vaccination Passports
Vaccination production data	Vaccination trials and reporting data	Vaccination “Yellow Card”	Cross-system data sharing
PPE testing results	Data used to inform public perceptions	No coordination across international boundaries – incompatible systems	Use of blockchain to validate Covid and vaccination status

This page is intentionally blank

9.3 Falsification / Misinformation of Data

One serious and perhaps unexpected aspect of the pandemic is that of misinformation. There are people either in denial of the virus's dangers, refusing to socially distance or refusing vaccinations. Reasons for these behaviours have generally been driven by intentional misinformation, ignorance, superstition, or economics. Marianna Spring, the BBC's specialist reporter covering disinformation and social media put the problem of misinformation succinctly when she stated "The problem with misinformation is that it is popular." See Barack Obama: One election won't stop US 'truth decay' – BBC News at https://www.bbc.co.uk/news/election-us-2020-54910344. Methods need to be devised or improved to prevent this effect, and to restore trust in carefully managed data.

This page is intentionally blank

9.4 Rumsfeld's known unknown and unknown unknown data conundrum

It is clear that until China reported to the world that Covid 19 had emerged as a threat, that data about the virus was an unknown unknown. However, we know there are thousands of viruses in animals that could pose a threat. These all need analysis and it may be possible to use massive computer analysis of genetic data to identify likely new threats.

A paper was given by Nick Hales as part of the 2021 Safety-Critical Systems Symposium which gives more detail on this topic *"Data Safety in Virus Outbreaks – Lessons learnt and Recommendations"* [**?**].

This page is intentionally blank

9.5 Learning

It is important to learn from these deficiencies because, while Covid 19 has brought tragedies with it, it is unlikely to be the last, or indeed, the most dangerous virus we will face. We must do better next time.

This page is intentionally blank

Acronyms, Definitions and Glossary (Discursive)

> *The plural of anecdote is not data.*
> ***Mark Berkoff***

This page is intentionally blank

9.6 Acronyms

This document is incomplete. The external file associated with the glossary 'acronym' (which should be called `Vol3.acr`) hasn't been created.

Check the contents of the file `Vol3.acn`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "Vol3"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "Vol3"
  ```

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.

This page is intentionally blank

9.7 Definitions and Glossary

This document is incomplete. The external file associated with the glossary 'main' (which should be called `Vol3.gls`) hasn't been created.

Check the contents of the file `Vol3.glo`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

If you don't want this glossary, add `nomain` to your package option list when you load `glossaries-extra.sty`. For example:

```
\usepackage[nomain]{glossaries-extra}
```

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "Vol3"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "Vol3"
  ```

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.

The normative list of definitions is at chapter 1. Normative definitions have been repeated here for convenience.

This page is intentionally blank

References (Discursive)

> *Data opportunities multiply as the data is transformed.*
> ***Sun Tzu misquoted***

This page is intentionally blank

Acknowledgements (Discursive)

> *Our ability to do great things with data will make a real difference in every aspect of our lives.*
> ***Jennifer Pahlka***

The document contributors would like to thank:

- The SCSC.
- The SCSC Covid-19 Working Group for providing some of the data used in the Covid-19 Appendix.
- Brian Jepson of the SCSC for web hosting support and technical help with the SCSC web site.
- Tim Rowe for editing this edition.
- Paul Hampton and Mark Templeton for managing the publication processes.
- Nick Hales, Mike Parsons, Tim Rowe, Alan Simpson and Mark Templeton for developing the additional text for this edition.
- Martin Atkins and Divya Atkins for driving the development of tooling and promoting data safety.
- Mike Parsons for chairing the Working Group meetings.
- All those who have taken minutes at Working Group meetings.
- All the organisations that have hosted Working Group meetings.
- All the organisations that have provided support to the document's contributors.
- Those that have been unable to attend meetings but have made supporting contributions.

This page is intentionally blank

Contributors (Discursive)

> *Without data, you're just another person with an opinion.*
> ***W. Edwards Deming***

This document has had the benefit of contributions from a large number of people, who work for a variety of organisations, which collectively span a range of different sectors. Note that contributions have been made on an individual basis and, in particular, the inclusion of an organisation in the following list does **not** necessarily mean that organisation agrees with the entire contents of the document.

Updates to the most recent version of the document were written by:

- Divya Atkins, Mission Critical Applications
- Martin Atkins, Mission Critical Applications
- Paul Hampton, CGI IT UK Ltd
- Mike Parsons, Ebeni and SCSC
- Tim Rowe, TGR Safety Management Ltd

In addition to the above, contributors to earlier versions upon which this document is based include the following (the organisations listed were correct at the time of their contribution) :

- Mike Ainsworth, Ricardo
- Rob Ashmore, Dstl
- Michael Aspaturian, EDF Energy
- Janette Baldwin, Thales UK
- Dave Banham, Blackberry QNX
- Ian Bingham
- John Bragg, MBDA UK Ltd
- Jennifer Brain, Wood plc
- Eric Bridgstock
- Simon Brown, QinetiQ
- Dermot Martin Burke, BAE Systems

- Dale Callicott, DKCSC Ltd
- John Carter, General Dynamics
- Martyn Clarke, SCSS Ltd
- Steve Clugston, TSC
- Robin Cook, Thales
- Davin Crowley-Sweet, Highways England
- Dijesh Das, AMEC / BAE Systems
- Duncan Dowling, DARD
- Andrew Eaton
- Ashraf El-Shanawany, CRA Risk Analysis
- Paul Ensor, Boeing
- Alastair Faulkner, Abbeymeade
- Ken Frazer, KAF
- Richard Garrett, SQEP
- Paulo Giuliani
- Ian Glazebrook, Atkins
- Rob Green, NATS
- Nick Hales
- Louise Harney, Leonardo
- Ali Hessami, Vega Systems
- David Higgins
- Gordon Hurwitz, Thales
- Pete Hutchison, RPS
- Gavin Jones
- Amira Kawar, Kawar Engineering Consultancy Ltd
- Tim Kelly
- Andrew Kent
- Brent Kimberley, Durham, Canada
- Julian Lockett, Frazer-Nash Consultancy Ltd
- David Lund, David Lund Consultants
- Dave Lunn, Thales UK
- Nasser Al Malki, University of York

- Victor Malysz, Rolls-Royce
- Jim Mateer, SQEP
- John McDermid, University of York
- Paul McKernan, Dstl
- Thor Myklebust, Sintef
- Mark Nicholson, University of York
- Yvonne Oakshott
- Robert Oates
- David Perrin, Virtual PV
- Ashley Price, Raytheon UK
- Andrew Rankine
- Felix Redmill, SCSC
- Sam Robinson, EDF Energy
- Mark Simmonite, Highways England
- Alan Simpson, Ebeni
- Oscar Slotosch, Validas AG
- Dave Smith, Frazer-Nash Consultancy Ltd
- Peter Smith, Highways England
- John Spriggs, NATS
- Carolyn Stockton, BAE Systems
- Mark Templeton, Arcade Experts Ltd
- Andy Williams
- Lesley Winsborrow
- Fan Ye, ESC

This page is intentionally blank

Index

DATA IS HERE. DATA IS GROWING. DATA IS CAUSING HARM.

This book has been developed by the Safety-Critical Systems Club Data Safety Initiative Working Group (DSIWG) to provide guidance on how data, as distinct from hardware and software can be managed in a safety-related context.

"If you torture the data long enough, they confess – even to crimes that were never committed."

Nihat Bülent Gültekin

~

This is the seventh minor update since version 3.0. Paragraph numbering within the body of the document remains aligned with that major release. Thus users of any previous 3.x release of the guidance document will find migration to this edition takes little effort.

SCSC

For Everyone Working in System Safety