ISSN 2754-1118
Vol 3, Issue 1

The Safety-Critical Systems Club

SAFETY-CRITICAL SYSTEMS eJOURNAL

SCSC
FOR EVERYONE WORKING IN SYSTEM SAFETY

ISSN 2754-1118 Vol.3 Iss.1

Editorial to the 2024 Winter Issue

Welcome to the first issue of the third volume of the Safety-Critical Systems eJournal, published by the Safety-Critical Systems Club (SCSC) with Publication Number SCSC-191. This is a themed issue, published to coincide with the Safety-Critical Systems Symposium, SSS'24. The theme is the technologies underpinning autonomous vehicles and how we assure them. The original brief was broad: *for example "technologies", as well as vision processing, machine learning, etc., can include things like concepts of operation, regulation, standards, ownership of liability, and so on*. Note that autonomous systems are influenced by human factors too.

When preparing an issue of the journal, it is expected that delays will occur due to illness or busy times for one or two of the authors and/or reviewers. This time the schedule was disrupted by the majority being ill or busy when needed; I hope everyone is fully recovered now. The plan to have three issues this year, as proposed in the previous issue, has had to be shelved; there will be two, but with more papers. Note that one of the papers originally scheduled for this issue has had to be delayed to the next issue; in its place we have a paper delayed from the previous issue…

This issue contains four papers:

- Philip Koopman and William H. Widen (USA) address what may be "safe enough" for automated vehicle technology in "*Breaking the Tyranny of Net Risk Metrics for Automated Vehicle Safety*". They consider legal, regulatory, ethical and equity aspects as well as risk metrics and industry standards conformance. This paper is the basis of a keynote speech on Day 3 of SSS'24.
- Peter Ladkin (Germany) applies Conceptual Analysis to electrotechnical terminology in "*Principles of Conceptual Analysis for Electrotechnical Terminology (ConcAn)*". He develops some principles and applies them to definitions from the International Electrotechnical Vocabulary, IEC 60050 (www.electropedia.org). This is a companion paper to that on semantic analysis, SemAn, in Issue 1 of Volume 2.
- Michael Wagner (USA) and Carmen Carlan (Germany) present "*The Open Autonomy Safety Case Framework"*. Developed over several years of work with the autonomous vehicle industry, their framework brings strategies, argument templates, and guidance together to support the development of a safety case for autonomous vehicle deployment.
- Jonathan (Jon) Wiggins (UK), in "*Human Factors in Functional Safety Assessment — Assessment of Human Interactions and Behaviour*", proposes a method to assess the impact of human factors and behaviour upon design, execution, and maintenance of a Functionally Safe System; it is presented as a framework that will be tailored for specific applications.

My thanks go to the authors for contributing their papers, and also to the anonymous peer-reviewers (at least three per paper) for suggesting improvements. Apologies also to those reviewers who made some recommendations that were not taken up.

The new cover image for this year includes logic, emergent properties, and machine learning motifs to mark the launch of the Club's new Working Group, Safe AI. Joining instructions will be announced at SSS'24. If you would like to find out more about this group, please go to their page on the SCSC website: https://scsc.uk/gb

John Spriggs, SCSC Journal Editor
February 2024

This collation page left blank intentionally.

ISSN 2754-1118 (Online)
ISSN 2753-6599 (Print)

Breaking the Tyranny of Net Risk Metrics for Automated Vehicle Safety

Philip Koopman[1] and William H. Widen[2]

1. Carnegie Mellon University, Pittsburgh PA, USA
2. University of Miami School of Law, Miami FL, USA

Abstract

An inquiry into how safe might be "safe enough" for automated vehicle technology must go far beyond the superficial "safer than a human driver" metric to yield an answer that will be workable in practice. Issues include the complexities of creating a like-for-like human driver baseline for comparison, avoiding risk transfer despite net risk reduction, avoiding negligent computer driver behaviour, conforming to industry consensus safety standards as a basis to justify predictions of net safety improvement, avoiding regulatory problems with unreasonably dangerous specific features despite improved net safety, and avoiding problematic ethical and equity outcomes. In this paper we explore how addressing these topics holistically will create a more robust framework for establishing acceptable automated vehicle safety.

1 Introduction

1.1 Overview

The obvious answer to how safe automated vehicles (AVs) should be is "net at least as good as a human driver". The AV industry typically argues that an expectation of a national reduction in net risk of fatalities warrants deployment of self-driving cars on public roads. This narrative dominates policy decisions, public messaging, and other aspects of stakeholder discussions of safety.

We argue that this focus on a net fatality risk metric is counterproductive to long-term success of the technology because that metric cannot be measured accurately at the early stages of deployment. Moreover, additional considerations will impact broad-based stakeholder acceptance. Reduction in net harm is a highly desirable goal, but over-overemphasizing this metric will likely back-fire on the industry due to neglecting other crucial metrics, ultimately contributing to a loss of trust in the technology. To meet with public approval in both the short and long terms, the AV industry must break free from the tyranny of a narrow net-risk metric approach.

In this paper, we explain that acceptably safe AVs must satisfy criteria along multiple different dimensions simultaneously. These criteria include:

- Achieving a Positive Risk Balance (PRB) for comparable conditions
- Mitigating risk transfer onto vulnerable populations

- Avoiding negligent computer driver behaviour
- Conforming to industry consensus safety standards
- Meeting regulatory requirements for risk on a fine-grain basis
- Addressing ethical and equity concerns

1.2 Previous Work

The topic of "how safe is safe enough" for autonomous vehicles remains a continuing discussion.

The BMVI (2017) Ethics Commission Report performed early influential work, creating a set of ethical rules for automated and connected vehicular traffic. Specific contributions included: requiring a positive balance of risks (no worse than human drivers), establishing a responsibility for regulators to ensure safety, minimizing risk to vulnerable road users, prohibiting decision logic regarding which specific road users to harm in an unavoidable crash, imposing responsibility on manufacturers for automated driving system safety rather than human vehicle occupants/drivers, and prioritizing preservation of human life vs. damage to animals and property. The BMVI report also addressed security and privacy concerns, along with other important considerations such as a need for people to retain some measure of control over equipment operation in appropriate circumstances. Our current work builds on that prior work, refining principles based on a half-decade of experience during which deployment of the technology has expanded.

The European Commission (2020) report on Ethics of Connected and Automated Vehicles, which we shall refer to as the "EC Ethics Report" sets forth twenty recommendations that cover not only safety, but also other issues such as informational privacy. It is a policy-oriented document which covers some of the same ground as our work, with sections on road safety risk, data/algorithm ethics, and aspects of responsibility. Its discussions are generally compatible with our work. We emphasize identifying different aspects of risk and safety that should be addressed in creating "safe enough" criteria that in some cases go beyond the scope of the EC Ethics Report, such as responder role contributions to safety.

A recent book focuses on this topic (Koopman 2022), emphasizing technical aspects of interest to a manufacturer wishing to determine that deployment is acceptably safe. Here we extend the scope to incorporate a broader range of concerns, including legal ones.

The SASWG (2022) published safety assurance objectives for all autonomous systems, generally emphasizing technical system properties in a safety engineering context, concentrating on computation, autonomy architecture, and system platform. That SASWG document also contains appendices with extensive cross-comparisons across other previous work to which we refer the reader. The SASWG work and its sources emphasize technical safety engineering considerations (e.g. how to ensure a predetermined level of safety). Here we add regulatory, legal, and a broader range of ethical issues into the scope of considerations.

The UK Centre for Data Ethics and Innovation published a study on the topic of Responsible Innovation for Self-Driving Vehicles (CDEI 2022), emphasizing legal and regulatory frameworks. That document does not directly present a framework for deciding whether a particular vehicle is safe enough for public road operation, though it does identify significant policy issues.

Chapter 5 of a Law Commission (2022) report proposes two paths for authorization for an AV to operate on public roads. One is Type Approval per international (UNECE[1]) standards. Another requires assurance of "at least an equivalent level of safety and environmental protection". Chapter 4 of that report additionally recommends avoiding risk transfers to vulnerable groups, which we also discuss in this paper.

Burton et al. (2020) cover a range of assurance considerations including engineering, ethical and legal, identifying a series of gaps: semantic, responsibility, and liability. They propose a framework including a safety case, dynamic assurance, soft law, and regulations that we believe is compatible with our findings, but takes a somewhat different approach.

This paper continues by covering each identified aspect of acceptable safety. It concludes with a composite statement of the relevant factors that should be accounted for in setting criteria for an acceptably safe autonomous vehicle deployment. We identify additional previous work specific to individual topic areas in the corresponding sections.

1.3 Terminology and Legal Framework

Here is a summary of some key technical terms and abbreviations used in this paper.

Automated Vehicle (AV): a vehicle with a computer driver which can completely carry out the driving task. A vehicle which has no requirement at all for a human driver might be called an "autonomous vehicle". However, in this paper we consider automated vehicles regardless of whether there is a human driver present, and focus solely on the safety of the computer driving function.

Computer Driver: a computer which controls steering and other aspects of motion control for a vehicle on public roads. A computer driver might or might not have a human backup driver who is monitoring operation, but who is not exercising sustained control of steering. (We sometimes use the term "AV" to refer to the behaviour of the vehicle as controlled by the computer driver to improve readability.)

Negligence: behaviour that fails to meet the level of care that someone of ordinary prudence would have exercised under the same circumstances (LII 2023).

Positive Risk Balance (PRB): the proposition that a computer driver should be no less safe (and ideally safer than) a human driver.

This paper is written based on the authors' knowledge of US laws. We understand that laws in other Western countries generally employ analogous principles. While some aspects of non-US law are addressed, that treatment should not be considered comprehensive. The treatment of legal issues is at level of abstraction such that differences should not alter our conclusions.

[1] The United Nations Economic Commission for Europe

2 Positive Risk Balance

2.1 Background

The notion of Positive Risk Balance (PRB) was a key contribution of the BMVI Ethics Commission report (BMVI 2017), captured in its rule number 2. While it notes that the long-term objective is to completely prevent harm to people, the report states that the "licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving, in other words a positive balance of risks". Moreover, BMVI's rule number 3 states that technologically unavoidable residual risks can be acceptable so long as PRB is achieved.

Since that time, the German car industry in particular has used PRB as its guiding star for acceptable safety.

2.2 How Much Better than a Human Driver?

While it seems intuitive that AVs should be at least as safe as human drivers, that is not the only possible approach. A long-view utilitarian argument might hold that the expected eventual reduction in fatalities from reduced road deaths morally justifies an increased fatality rate in the short term to speed up the development of a safety panacea. The AV industry might be said to obliquely take that approach via its efforts to minimize regulatory oversight and speed deployment based on arguing there is a moral imperative to reduce road deaths. Such a narrative glosses over the ethical problems with potentially near-term increasing risk to road users while hoping technology matures to provide promised benefits. A testing fatality in 2018 (NTSB 2019), well before one might have expected such an event based on human driver fatality rates, illustrates that dynamic, and justifies concerns over elevated near-term risks due to hurried development efforts.

Some data suggest that a modest PRB might not be sufficient. Liu et al. (2018) found that while a risk of 4 to 5 times as safe as a human driver might be tolerable, a broadly acceptable risk goal would be a hundred-times improvement in safety. It remains unclear whether public stakeholders will accept only a modest net safety improvement. One report popularized the notion that 10% better than human drivers should dictate a decision to deploy the technology (Kalra and Groves 2017), essentially arguing that any measurable decrease in fatalities creates an imperative to deploy sooner rather than later.

2.3 Multi-dimensional PRB Comparisons

While a safety metric of "better than a human driver" sounds intuitively appealing, actually measuring such an outcome requires establishing a comparable baseline for a human driver.

> ***Hypothetical example:*** *A new AV with high-end active safety features being driven in fair weather on empty city streets is compared to a statistical average human driver baseline. That baseline human driver is operating a statistically average lower-trim vehicle 12 years of age, with less capable, outdated safety features. The baseline human driver is also operating in conditions that include dangerous secondary roads, twilight, snow, and with some aggregate fraction of impaired driving.*

Defining and characterizing a human driver baseline can be surprisingly complex, involving contributions from factors such as (Koopman 2022):

- Driving environment (city/urban, light/dark, dry/wet/snow/ice, road maintenance condition, class of roadway, prevalence of vulnerable road users, local driving customs, etc.)
- Vehicle type (weight, installed passive safety features, installed active safety features, maintenance condition, etc.)
- Driver demographics (driver age, driver experience, any driver impairment, any driver distraction, any violation of road rules by driver, etc.)
- Victim demographics (age, pedestrian/bicyclist/motorcyclist/driver/passenger, etc.)

These factors can have a dramatic impact on the baseline crash rate and expected fatalities for human drivers, often comparable to the excess safety factor of 4 or 5 times safer that might otherwise have been thought adequate. As a single example, roadway type accounts for approximately a factor of 5 difference in fatality rate per mile in Pennsylvania when comparing the safest roadway (the Pennsylvania Turnpike) to the most dangerous type (non-Interstate system highways) (PennDOT 2022, page 16).

To be a fair comparison, a PRB calculation would need to include a weighted average of contributions to harm from data accounting for various contributing factors. Some factors might be excluded for policy reasons, such as not including risk contributions from drunk drivers as part of a policy decision to make the baseline an unimpaired driver.

Creating a human driver baseline would require detailed data for human driver rates of harm in various combinations of circumstances. While this is technically achievable, it is far from a simple comparison of national average fatality rates. It is possible that a very large safety margin (an order of magnitude or more) could simplify the comparison process so that minor factors could be neglected in analysis, but only at the cost of demanding much higher AV safety performance than might be strictly necessary to achieve credible PRB.

2.4 PRB and Risk Subsidy

A more subtle issue has to do with the possibility of risk subsidy.

> ***Hypothetical example:*** *AVs have a 10% reduced net fatality rate, accounting for all relevant environmental and driver type PRB conditions, but not equipment type. Upon review, it becomes apparent that the entirety of this safety improvement is attributable to the installation of an Automatic Emergency Braking (AEB) feature on all AVs. In comparison, the older human-driving vehicle fleet that includes a high proportion of lower-cost vehicles has a comparatively low installation rate of AEB. Moreover, it is found that turning on an AV computer driver actually reduces safety compared to manual operation of that same vehicle in comparable conditions (but not enough to be as bad as human driver outcomes in the average vehicle fleet). The computer driver itself reduces safety, but this is masked by the AEB feature safety improvement when compared to a fleet of predominantly non-AEB-equipped vehicles.*

This type of scenario is one possible outcome for automating driving that requires continuous human driver supervision. Data analysis from Goodall (2023) shows that early adopters of an automation feature saw an 11% adjusted estimated crash rate increase with the automation feature on vs. off. However, net safety with the automation feature on was

still said to be much better than for the US vehicle fleet which, on average, lacked comparable active safety features.

We characterize such a situation as a risk subsidy: some active safety technology is added to increase net safety. Then a vehicle automation feature is added that increases total harm substantively, but not enough to make the vehicle worse than it would be without that additional active safety technology. Public stakeholders might consider an AV safety claim based on a risk subsidy to be a misrepresentation of the safety benefits of AV features.

2.5 PRB as One of Many Metrics

Even if manufacturers go to the significant lengths required to implement a PRB metric with detailed weightings to account for the various factors involved, it is unlikely that this metric will show conclusive evidence of fatality reductions during initial deployments in the first several years of operation. The reason is a simple statistical significance issue. Current outcomes are approximately one fatality per 74 million miles driven (NHTSA 2023a), and one fatality per 192 million miles in the UK (UK DfT 2022). Given that at the time of this writing robotaxi companies are claiming perhaps one to three million miles each (Bidarian 2023), there might be two more orders of magnitude mileage accumulation to get a meaningful understanding of fatality outcomes. Accounting for miles driven by different ADS models remains problematic, as does accounting for accumulation of miles across multiple software upgrades.

Because property damage and injury crashes are much more prevalent than fatalities in human driver data, statistical confidence as to PRB outcomes for those less severe instances of harm will come earlier. However, it remains to be seen how accurate fatality predictions will be based on less severe crash rates, even with a sophisticated “inverting the diamond” analysis and prediction approach such as that used by Waymo (Victor et al. 2023). A significant confounder for early predictions is the potential for common cause failures, such as a defective software update causing a cluster of high-severity crashes before the update could be rolled back.

Additionally, there are risks from large disruptions in the operational environment causing failures due to the general brittleness of machine learning technology to novel situations. A large power outage affecting traffic signals (Fleischer 2023), overloaded mobile data networks (Hawkins 2023), a bodged software update, or other adverse event that affects hundreds of cars simultaneously could potentially cause large numbers of severe common cause mishaps in ways that human drivers are unlikely to experience, potentially invalidating forecasting approaches based on an assumption of random independent failures. An overarching concern is that while putting the same computer driver in every vehicle provides a basis for fleet-wide learning and improvement, it also introduces a source of common cause failures across a deployed fleet.

Another potential issue with pure PRB approaches is that in practice they emphasize aggregate total risk, and aggregate total harm. The intuitive appeal is that if we can reduce total road fatalities by even 10% over a relevant human driver baseline, that will be an improvement in highway safety and should be considered a victory for adoption of AV technology. However, there are possible outcomes of net PRB that might still be societally unacceptable.

Hypothetical example: *Total fatalities are reduced by 50%, but every single person harmed is a child boarding or debarking a school bus. (Scenario inspired by a Tesla mishap (Krisher 2023)).*

Hypothetical example: *Total pedestrian fatalities are reduced by 90%, but every single fatality is due to an AV rolling through stop signs. (Scenario inspired by a Tesla recall (Gitlin 2022)).*

While perhaps too specific to be likely in practice, these hypothetical outcomes are intended to illustrate the point that there are aspects of socially acceptable safe outcome characteristics that go beyond pure aggregate PRB. As one researcher noted, "*An AV that kills 1,000 fewer car occupants but 100 additional pedestrians may not be acceptable, even though 900 net lives are saved*" (Goodall 2021).

PRB summary: Positive Risk Balance is a necessary but insufficient condition for safety outcomes that are likely to be broadly acceptable to a wide range of stakeholders. Other considerations such as the demographic profile of victims and whether loss events are associated with the AV violating road rules are likely to be relevant as well. Moreover, computing PRB is complicated because of the need for a comparison to a detailed human driver baseline that accounts for varied factors.

3 Risk Transfer

3.1 Preamble

Even if net harm is reduced, it might well be that transfer of risk from one population or demographic segment onto another is considered societally unacceptable. The EC Ethics Report (2020) addresses this with its recommendation 1, noting that no category of road user should be at risk of increased harm, even if net harm to all road users has been reduced. That report goes further, suggesting in its recommendation 5 that AVs should adapt their behaviour to redress existing risk inequalities.

3.2 Statistical Risk Transfer

Hypothetical example: *A computer driver is involved with a series of crashes at emergency response scenes, causing both injuries and fatalities. Public pressure forces the manufacturer to deploy a software update to address the issue even before data is available as to whether this is a worse risk than that presented by human drivers in similar situations. (Scenario inspired by Tesla investigation by NHTSA (Hawkins 2022).)*

A more general issue is that if net harm is decreased, but harm to a distinguishable group of road users is increased compared to a relevant baseline, that risk transfer seems unlikely to be acceptable to public safety stakeholders. This applies especially if the group seeing an increase in harm from AVs is considered particularly vulnerable. Examples of such population segments might include: road workers, people with sensory impairments, people with mobility restrictions, children, the elderly, pedestrians, bicyclists, wearers of distinctive

ethnic clothing styles, people with darker skin tones, and historically disadvantaged groups. This issue will apply not only to harm from operation, but also potential harm that might occur during public road testing of the technology (Widen 2022). A similar risk transfer occurs between present road users and future road users if present road users are exposed to an increased risk of harm on the expectation that the lessons learned from earlier deployments will increase safety for future road users — an ethically controversial deployment decision (Widen 2023). This situation also raises the question of the appropriate discount rate for anticipated future benefits, both for time value of money and likelihood of realization (Johnsson and Voorneveld 2018).

A related issue is if there is an identifiable pattern in loss events that might be possible to avoid with a design improvement. AV proponents argue that the perfect should not be the enemy of the good (Kalra and Groves 2017), meaning that the technology should be deployed as soon as there is a net statistical reduction in harm. However, it is predictable that systematic losses that might be mitigated at reasonable cost will produce adverse reactions from at least some public safety stakeholders — even if the net harm is reduced.

3.3 Intentional Risk Transfer

A different sort of risk transfer might happen intentionally based on design decisions. The classic example of this is the so-called Trolley Problem applied to AVs. In that version of the Trolley Problem, a computer driver is said to be presented with a binary no-win situation in which somebody will necessarily die (Koopman et al. 2021). The question is which victim(s) the computer driver should choose, often couched in terms of some multi-dimensional utilitarian calculation of the number of lives and the comparative value of each life involved. Burton et al. (2020) discuss why such a framework is an unhelpful analogy for AV safety. EC Ethics Report (2020) recommendation 6 is to proactively address such situations.

BMVI (2017) rejects the notion that a calculus of any sort should be used to offset one life against another in an unavoidable crash, especially in its rule number 9. On a per-crash basis, BMVI rule 9 also requires no transfer of harm onto those not directly benefiting from vehicle automation. For example, an innocent bystander pedestrian should not be sacrificed in hopes of avoiding a crash that would be fatal to multiple vehicle occupants. However, BMVI also admits that there might be some circumstances in which reducing the total number of innocents harmed in an unavoidable crash is justifiable, so this is still an open topic to some degree.

While the question of intentional risk transfer can raise serious moral questions, for the near term those incidents should be a small fraction of the total number of mishaps. (If this is not true, likely the computer driver has much more serious safety issues.) A relatively simple behavioural constraint might suffice pending further study of the problem.

> ***Hypothetical example:*** *An AV is designed so that it tries to avoid harming any person. However, once its software deems it reasonably likely to inflict harm on one or more people, its manoeuvring options are restricted to avoid harming any additional people who would not already be harmed by the existing trajectory plan.*

With this example design approach, the computer driver is permitted to act to reduce the severity or attempt to avoid harm, but is not permitted to intentionally harm any individual who is not already expected to be harmed when it becomes apparent that a loss event is unavoidable. This policy would, for example, prohibit swerving to avoid striking a tree if

that swerve would instead run over a single pedestrian — even if doing so would be likely to save multiple vehicle occupants from severe harm. While such a strategy is likely not optimal in at least some utilitarian sense, there is no general agreement on what "optimal" might really mean in practical situations. Until consensus is reached on a better strategy (if that is even possible), it is a defensible strategy aligned with BMVI guidelines (BMVI 2017) that is relatively straightforward to define and implement.

Risk transfer summary: AVs should minimize or eliminate risk transfer. This is likely a practical requirement for public safety and equity stakeholders, even if doing so arguably limits the potential safety benefit of the technology by requiring constrained loss mitigation strategies.

4 Lack of Negligent Driving Behaviour

4.1 Negligence

An additional limitation on AV safety is likely to be a prohibition against negligence. We define *negligence* in this context as driving behaviour which, if exhibited by a human driver for a relevant set of conditions, would be considered negligent (or reckless) according to applicable laws, statutes, ordinances, and regulations that might apply. The scope of negligence must include both tort law and criminal law. Simply put, this is holding computer drivers to the same standards of negligent and reckless driving behaviour that already apply to human drivers.

4.2 An example of Negligent Computer Driving

We do not here attempt a comprehensive treatment of legal issues. Rather, we use an example to illustrate relevant concerns.

> ***Hypothetical example:*** *An autonomous vehicle is driving on city streets. A licensed driver who initiated autonomous operation an hour earlier is present in the vehicle, but asleep in a reclined seat as provided by the manufacturer. ("Take a nap and leave the driving to us!") The computer driver runs a red light and strikes a pedestrian in a marked crosswalk, killing the pedestrian instantly. The police arrest the sleeping vehicle occupant for negligent homicide or a similar offence.*

In this example, if the human driver had been driving, that human driver likely would be found negligent in the absence of significant extenuating circumstances such as brake failure, and potentially charged with negligent homicide. This is due to the situation in which violation of a traffic law (running a red light) directly led to a fatality under a presumption of negligence *per se*.[2]

The question of how such a situation is handled when a computer is driving is unsettled. The authors have proposed that the computer driver should be held to the same standards as

[2] "A defendant who violates a statute or regulation without an excuse is automatically considered to have breached her duty of care and is therefore negligent as a matter of law" … "The most common application of negligence *per se* is traffic violations, where the driver is automatically considered negligent for violating the traffic code." Legal Information Institute. (2020 update). Definition of "*negligence per se*". Wex (Cornell). Available at: https://www.law.cornell.edu/wex/negligence_per_se

a human driver, with the manufacturer held to be the responsible party in such a situation (Widen and Koopman 2023a).

Regardless of how negligence might be handled legally, as a practical matter, it seems highly desirable that computer drivers exhibit vanishingly small amounts of negligent driving behaviour. Deploying vehicles that ignore traffic signals because the designers think that it is safe to do so per their design should not permitted. Computer drivers should obey traffic laws. If manufacturers find traffic laws overly restrictive, they should lobby governments to change the traffic laws instead of flouting them[3]. Regulators seem to agree with this statement. For example, Tesla vehicles were recalled for not coming to a full and complete stop at stop signs (NHTSA 2022a).

BMVI (2017) holds that product liability should be the governing principle for harm caused by computer drivers. While product liability should be a possible avenue for collecting compensation for harm, conventional fault-based tort law also should provide an avenue for recovery because it is a more efficient and cost-effective process for attributing liability in garden-variety accident situations in which a computer driver might have behaved in a clearly negligent manner.

This issue goes beyond arguing for a general respect for the rule of law. Tort law and criminal law do not recognize statistical safety arguments to determine the fact of culpability. To put it bluntly, someone who saves 1000 lives does not get a coupon for one free homicide. Someone who has a perfect driving record for 40 years does not get a free pass for causing a crash by running a red light while drunk. While a person's history and character might be weighed in determining an appropriate penalty, guilt is determined by the facts in the particular case, not the history of the individual. For example, a first time Driving Under the Influence offender may qualify for a diversion program unavailable for repeat offenders (Bieber 2023).

By the same token, a computer driver that is documented to cut fatality rates in half should not get a free pass on negligent driving in a particular case, unless the public policy approach is to pre-empt tort law and criminal liability for computer drivers. Even so, it seems unlikely that public stakeholders will welcome AVs that cause fatalities associated with egregious traffic rule violations that they perceive would not have been excused for a competent human driver, even if total fatalities are reduced.

4.3 Responder-Role Safety

Lack of negligent behaviour is a highly desirable aspect of safety, but seems unlikely to ensure acceptable safety on its own. In particular, blaming other drivers for easily preventable crashes is likely to degrade safety outcomes if blame for contributory negligence or comparative fault results in the exclusion of data from safety metrics.

> ***Example:*** *An autonomous vehicle is making an unprotected turn across oncoming traffic on a multi-lane road at night. There is an oncoming vehicle that is speeding. The AV calculates that the oncoming vehicle must slow down to make a turn onto a side street as required by a pavement "turn only" lane marking, and that this gives enough time for the AV to turn into that same street ahead of the oncoming car. The oncoming car does not slow down, and does not make the turn. The AV, sensing a failed plan, executes an emergency safety stop in the oncoming car's lane, with a subsequent*

[3] EC Ethics Report Rule 4 reasonably suggests considering revision of traffic rules and setting non-compliance policies for AVs.

collision. The AV company states that since the oncoming car was more at fault due to speeding and not making the required turn, the multiple injuries from the collision should not count against the AV's safety record. (Inspired by a robotaxi mishap (CA DMV 2022).)

This example involving both vehicles as contributors to the incident illustrates important aspects of the interaction of blame and safety.

If the other vehicle had not been speeding, more blame might have fallen on the AV instead. There is no indication in the crash report that the speeding was a factor in the computer driver's decision making. A crash still would likely have happened if similar timing issues had been involved with a non-speeding oncoming vehicle. This suggests that the decision to make the left turn in this case was unduly risky, based as it was on an assumption that another road user would manoeuvre despite lack of indication from that other vehicle of a manoeuvring intention[4].

Arguably the reaction of the AV to perform an immediate stop when it calculated that it might crash is problematic. There are circumstances in which making an immediate stop might increase risk rather than reduce it, such as making a sudden stop during an unprotected left turn with high-speed oncoming traffic. This illustrates that the action taken by a computer driver when there is a problem will sometimes need to be more than a simple emergency stop. Stopping in ways that block emergency responders provides an additional illustration of the problem with assuming that stopping is always a safe behaviour (Nicholson et al. 2023).

Finally, the AV would have been better off if it had calculated that there was likely uncertainty in the behaviour of an oncoming vehicle, and waited for that vehicle to pass before initiating the left turn. Occurring late at night as this mishap did, it is likely that a paucity of traffic gave plenty of opportunity to make the turn safely after that oncoming speeding car had passed.

A more general view of this topic is that the role of a responder matters for safety (Victor et al. 2023). Even though one vehicle might behave in a negligent manner (whether by a human or computer driver), harm might still be avoided by other responding vehicles taking evasive actions. For example, the computer driver might have determined that the speeding car might not slow down, and instead waited to make the left turn. In another scenario a responder might slow down to delay entering an intersection when it detects cross traffic is likely to proceed through a red traffic signal. In practice, responders avoiding crashes that might otherwise be blamed on negligent driving by the other driver can make a substantial contribution to road safety.

From a public stakeholder perspective, it is likely unacceptable for AVs to make seemingly "stupid" driving decisions (based on what public perception of a reasonable human driver behaviour might be) such as turning in front of an oncoming speeding car, and then attempting to claim innocence because the other driver was found more than 50% at fault, despite a substantive fraction of blame being assigned to the AV in a crash investigation.

This has three practical implications for setting acceptable safety characteristics:

[4] We note that the October 2, 2023 loss event in San Francisco which resulted in the suspension of activities by Cruise LLC may present an issue of comparative fault. We do not use that example because, at the time of this writing, the facts of that case remain under investigation.

1. Computer drivers should have robust skills in the role of an incident responder. This roughly corresponds to having good defensive driving skills.
2. As a more specific part of defensive driving behaviours, computer drivers should have robust models of other road user potential behaviours that includes other vehicles violating traffic rules in readily foreseeable ways (speeding, rolling stops, using an incorrect lane, entering an intersection shortly after their traffic signal turns red) and pedestrians not using official crosswalks, which contributed to the Uber ATG pedestrian fatality (NTSB 2019).
3. Risk analysis of safety manoeuvres such as in-lane stops should account for potential harm that might occur after the stop, such as being hit by other road vehicles or trains, and disrupting emergency response services.

4.4 The Role of Blame in Safety

Care must be exercised when invoking the notion of blame in setting acceptable risk criteria. Lack of legal fault for AV behaviour does not equal an acceptable safety outcome when AV behaviour is a contributing factor to an accident, even if not the primary factor.

> ***Hypothetical example:*** *A robotaxi fleet is hit from behind frequently, each time blaming the trailing driver for the crash. The net result is, however, that the robotaxis are involved in twice as many rear-end crashes as a human driven baseline, even though not a single crash is blamed on the AV. (Scenario inspired by (Stewart 2018).)*

Computer drivers being hit from behind at low speeds is commonly attributed to driving behaviour deviating from the expectations of a trailing human driver, sometimes being characterized as the AV being overly cautious in comparison to prevailing human driving norms. A commonly cited reason for such crashes is the AV displaying so-called "phantom braking" behaviour in which an AV panic stops for no reason that is discernible to trailing vehicles. While the blame by default is imposed on the trailing car in the crash, blame assignment on its own seems unlikely to improve a trend in rear-end crashes in those circumstances.

While it is desirable to avoid at-fault crashes for AVs, safety metrics should not exclude crashes that are not the AV's fault as determined by legal standards. Rather, they should show that both (a) at-fault crashes are better than an ordinary at-fault human driver baseline, and (b) total crashes (both at-fault and not-at-fault) are also better than an overall human driver baseline. Proportional fault should not be used to claim that an AV was not at fault unless the proportion of fault assigned is negligible (perhaps less than a few percent), and not merely 50% or less fault as found in some comparative fault systems[5].

The EC Ethics Report (2020) recommendation 19 goes further, recommending a fair system for attribution of moral and legal culpability. That report's recommendation 20 is to establish a fair and effective mechanism for granting compensation to those harmed by an AV.

Negligence and Blame Summary: AVs should minimize the rate of crashes caused by computer driver behaviour that would be considered negligent it performed instead by a human driver. In particular, an improvement of net statistical safety should not be used as

[5] A pure comparative fault system considers all fault from whatever source and in whatever percentage. A modified comparative fault system does not allow recovery if the plaintiff is 50% or more at fault. A contributory negligence system allows no recovery if the plaintiff is found to have any fault.

an excuse to forgive negligent computer driver behaviour. Net safety metrics must include all crashes, not just those in which blame has been assigned to the computer driver, to encompass the contribution of defensive driving skills to net safety outcomes.

5 Standards Conformance

A significant challenge in defining acceptable AV safety is validating the accuracy of leading indicators (Kalra and Groves 2017) to predict safety outcomes that might not be statistically measurable until many years in the future. Sophisticated predictive approaches based on human-centric safety improvement techniques can be applied to provide improved confidence. However, in the final analysis there are assumptions and threats to validity to any predictive technique. Net risk prediction accuracy, especially for fatalities, will not be established until hundreds of millions of miles of real-world road usage have been accumulated.

The AV industry as a whole has been focussed on public road testing as a way to predict eventual safety[6]. However, in other industries a primary method of assuring safety deployment is following industry consensus standards. It is remarkable that the automotive industry, in sharp contrast to other life-critical technology industries, has historically not been required to follow its own consensus safety standards to deploy on US public roads[7].

One way to improve confidence in predictions of eventually acceptable safety would be for AV manufacturers to follow industry consensus standards. While the list of potentially relevant standards is large and growing, some key international standard candidates include: ISO 26262:2018, ISO 21448:2022, ANSI/UL 4600:2023, ISO/SAE 21434:2021, and SAE J3018_202012. Additionally, a Safety Management System should be in place, perhaps structured using advice from the Automated Vehicle Safety Consortium (AVSC 2021).

A typical industry talking point against standards is that they are said to "stifle innovation." This is the sort of thing that tends to be said by those who see safety practices as an impediment to risky innovation, such as the creator of the Titan mini-sub who perished in an implosion event (Musumeci and Guenot 2023), or companies simply looking for an excuse to evade regulatory oversight.

Standards tend to be written in the metaphorical blood of past mishaps. Moreover, the standards referred to above do not constrain the specific technology used to implement AVs. Rather, they require the use of hazard and risk analysis approaches with accompanying mitigation approaches. Those standards also require accounting for known issues with life critical systems such as risks posed by common cause failures, and encourage addressing problems that are foreseeable enough that they have been included in an international standard.

Standards Summary: There is no need for the AV industry to relearn lessons the hard way via loss events that could be avoided. Standards conformance is a key technique other industries use to establish a justifiable belief in acceptable safety before deployment.

[6] The EC Ethics Report (2020) points out in its recommendation 3 that road testing introduces its own risks.

[7] The authors are not aware of an ISO 26262 conformance requirement for any country for conventional road vehicles. It seems that ISO 26262 is strongly encouraged in Germany for fully automated vehicles, e.g. per Appendix 1-Autonomous Vehicles Approval and Operation Ordinance (AFGBV), Part 1, 1.3 Planning of routes and speeds; 7.2.1 Hazard Analysis Available at: https://www.buzer.de/Anlage_1_AFGBV.htm?m=26262#hit. The authors thank Gabi Escuela for bringing this to our attention.

6 Regulatory Requirements

6.1 Safe Enough?

Any AV will additionally need to meet regulatory requirements specific to the country or region in which it is being deployed. The rules differ by region, but conformance to relevant rules is a required part of the characterization of acceptable safety.

Regulators also need a criterion by which to decide what might be "safe enough", which is commonly expressed in terms of an acceptable risk threshold. In the US and the EU, the prevailing criterion is "Absence of Unreasonable Risk" (AUR), whereas in the UK the primary criterion is "As Low As Reasonably Practicable" (ALARP).

6.2 Testing-centric Requirements

Testing-centric requirements base a compliance determination on a test that can be replicated by an independent party on a series production vehicle. In the US, the Federal Motor Vehicle Safety Standards (FMVSS) serve this role (NHTSA 2023b).

The FMVSS suite requires specific safety-related functionality and safety feature performance. Topics range from tire pressure warning indicators to rear-view cameras to passenger crash safety features and more. FMVSS compliance is self-certified by manufacturers, and subject to audits by regulators after deployment. There is no pre-release regulatory approval process in the US for ground vehicles.

Of particular concern for AVs is that the FMVSS collection was created with many test procedures explicitly requiring the presence of a driver seat, steering wheel, and other equipment which might not be present in a cargo AV or completely driverless passenger AV. While the FMVSS 200 series on crash safety has been modified to address this constraint (NHTSA 2022b), work continues on other portions of FMVSS to remove inapplicable dependencies on human driver support equipment. A topic of intense political lobbying by the industry is increasing the number of FMVSS exemptions to permit scaled-up deployment until those other aspects of FMVSS can be addressed (Congress 2023). Granting a large number of exemptions can amount to deregulation, and not merely a deviation for small-scale controlled testing.

Europe has a type approval approach, with acceptance tests performed by independent parties under contract to manufacturers in a process known as homologation. That process results in a government-issued certificate granting permission to sell a particular vehicle (European Commission n.d.).

An additional aspect of testing-centric requirements is the NCAP[8] process, which involves a star rating customer information approach rather than a hard regulatory requirement for a particular level of performance. The theory is that consumers will favour vehicles with higher safety ratings disclosed at the time of sale. The US has come under criticism for its NCAP system lagging behind EURO-NCAP in adopting tests relevant to automated driving features (NTSB 2022).

It is possible for test-based regulations to extend to vehicle automation features, which was the case with the UN ECE #157 regulation on Automated Lane Keeping Systems (ALKS)

[8] New Car Assessment Program

(United Nations 2021). This is a regulatory approach for European approval of low-speed highway traffic jam pilot vehicle features.

A particular challenge to all types of regulatory approaches is the recent increase in frequency and safety relevance of over-the-air software updates that change a vehicle's software via a remote software update rather than a dealer visit. These are increasingly used not only to deliver remedies for safety recalls, but also to deploy new features which might introduce safety problems of their own. Regulators are still struggling to address over-the-air updates within their regulatory frameworks (Stumpf 2021).

6.3 Risk Reduction

Another concept used in regulations that can put a constraint on the acceptable boundaries of AV performance is that unreasonable risk should not be present, with a prevalent threshold being an Absence of Unreasonable Risk (AUR). AUR is a concept determined by a multi-point set of criteria in the US including: the utility of the product, the level of exposure to the risk, the nature and severity of hazards, and the likelihood of resulting harm. Also relevant are the state of the art, the availability of alternate designs, and the feasibility of eliminating the risk (US CFR 1992).

The US regulator — the National Highway Traffic Safety Administration (NHTSA) — uses AUR as a cause for pursuing regulatory enforcement action in addition to failure to comply with FMVSS (NHTSA n.d.).

In practice, AUR operates primarily at the feature level, not the vehicle level. The main mechanism available to NHTSA is a recall, which is a regulatory action taken only after a particular vehicle feature has been deployed on public roads. It is commonly the case that a recall is only pursued after a substantial number of incident complaints have been filed, and sometimes only after multiple reports of harm. NHTSA is, however, quietly experimenting with a regulatory strategy requiring immediate crash reporting and directing rapid recalls to change defective ADS software (Wansley 2022).

While AUR might be used as an overall design goal by manufacturers, in practice the application of AUR by regulators tends to be recalls for specific, documented dangerous behaviours or design deficiencies. It would require a dramatic expansion of their historical regulatory approach to ban an entire vehicle based on a failure to achieve net PRB or the like based on overall crash rates. While one initial somewhat broader recall has been promulgated (NHTSA 2023c), this seems to be breaking new ground for regulators.

In the UK, and some other countries, a general regulatory approach requires risks to be As Low As Reasonably Practicable (ALARP) or the equivalent (HSE n.d.).

At a high level of abstraction, ALARP has more similarities than might be apparent with an AUR approach. In implementation, the question comes down the how the cost/benefit decision is made in terms of how much risk reduction is "reasonable" or "reasonably practicable" for a particular AV in its expected operational environment. The regulatory emphasis in practice in both cases ends up being whether any particular risk or specific potential design defect could have or should have been mitigated further, rather than the overall statistical rate of harm.

Regulatory Summary: Regulations provide a combination of vehicle-level testing requirements and retrospective requirements to mitigate risks that emerge after deployment.

Neither of these are directly aligned with a PRB metric, and so create additional requirements to achieve acceptable safety.

7 Ethical and Equity Concerns

7.1 Preamble

While it is typical for a discussion of acceptable safety levels to emphasize technical and regulatory requirements, from a public acceptance point of view ethical and equity issues are also important. We consider the "Moral Crumple Zone" issue, broader ethical imperatives, and equity considerations that should factor in to ensuring acceptable safety.

7.2 Moral Crumple Zone

The concept of a Moral Crumple Zone is a design strategy to shield a technological system from blame at the expense of the nearest convenient human, regardless of whether it is reasonable to expect that human to have been able to avoid or mitigate loss events (Elish 2019). It deflects attention from system shortcomings which place humans in unrecoverable situations or otherwise set them up to fail.

The classical example of a moral crumple zone in the AV domain is the story of the tragic Uber ATG fatality.

> ***Real example:*** *An AV test vehicle strikes and kills a pedestrian at night. The driver is said to have been distracted by a mobile phone at the time of the crash, and did not notice the pedestrian in time to stop, even though adequate sight line and time were available. The technical cause of the mishap was a combination of tracking software defects, an expectation that pedestrians would only cross at official crosswalks, and the disabling of a manufacturer AEB system. A National Transportation Safety Board (NTSB) investigation found that the driver was not paying attention, but also found that profound deficiencies in safety culture at Uber ATG set the stage for the mishap. The company, Uber ATG, settled with survivors but was not charged criminally. The test driver was sentenced to supervised probation after a plea deal involving a felony endangerment charge (NTSB 2019) (Smiley 2023).*

A significant issue with the Uber ATG fatality is that one can argue the safety drivers were set up for failure due to lack of supervision, onerous working conditions, immature technology, and the inevitability of automation complacency. It seems likely a matter of when, and to which test driver — rather than if — such a crash would happen. While it is reasonable to hold test drivers accountable for a failure to diligently perform their duties, it does not serve the interests of safety to let companies use those same test drivers as a defensive shield against accountability for dangerously run test programs.

In the wake of the Uber ATG fatality, the industry updated the SAE J3018_202012 test driving safety standard to incorporate lessons learned. However, no currently operational company publicly states that it conforms to that industry consensus standard[9].

[9] Argo AI was independently assessed to conform to SAE J3018 (PRNewswire 2021). However they have recently terminated operations.

A related issue is Tesla's use of retail customers as so-called "beta testers" for obviously immature automation technology. Tesla has succeeded in laying the blame for crashes in court on drivers for failing to mitigate driving errors made by the technology in an initial case (Roy et al. 2023). A criminal case similarly saw the driver take responsibility for a fatality that occurred while Autopilot was in use as part of a plea agreement (Dazio and Krisher 2023). It is unclear whether this trend of blaming drivers will continue despite concerns of the NTSB regarding automation complacency for this technology (NTSB 2019).

7.3 Ethical and Equity Considerations

There are many other ethical and equity considerations that should be incorporated as appropriate into acceptability criteria for a safety program. Some of them might be considered to stray a bit far afield from traditional functional and system safety. However, they can directly impinge upon and constrain design choices, activities, and deployment decisions that are relevant to safety. Examples include:

- Whether developmental testing of potentially defective prototype vehicles is allowed in historically disadvantaged areas, where it can be argued that residents harmed are likely to receive lower compensation from defendants than would residents of other areas (Widen 2022).
- Whether states or municipalities should be pre-empted by higher layers of government from instituting limitations on testing or deployment of AVs responsive to specific hazards in their local communities, such as a prohibition on testing in sensitive locations such as active school zones (Widen and Koopman 2022).
- The level of public transparency afforded to mishap reports from manufacturers that are required by regulators. Current national-level mishap reports (NHTSA 2023d) have substantive redactions of arguably non-technical data such as crash locations in the name of manufacturer proprietary data secrecy.
- Whether an argument that an aspirational goal of eventual reduction in harm should be permissible as a justification for deploying immature technology on public roads when it is not known when (or even if) PRB and other acceptable safety criteria will be fulfilled.
- Who, if anyone, should be held accountable for negligent driving behaviour on the part of a computer driver. Current US laws are anything but clear on this topic (Widen and Koopman 2023b).
- The degree to which disruption of public safety functions and emergency responders (Nicholson et al. 2023) should be factored into PRB baselines and other safety metrics.

Conformance to the IEEE 7000-2021 standard might provide a way to identify and address ethical considerations from a wide range of stakeholders, but is not currently required by any jurisdiction of which the authors are aware.

Ethics and Equity Summary: A number of questions regarding ethics and equity remain open, but should be addressed by any proposed set of criteria for acceptable safety. A particular area of practical concern is avoiding the use of test drivers or retail customer drivers as Moral Crumple Zones to shield manufacturers from accountability for potentially defective, immature driving automation features.

8 Summary

Recapping the section summaries, acceptable safety criteria for a computer driver should encompass all of the following areas:

- Positive Risk Balance using a baseline that accounts operating conditions and vehicle safety features like-for-like.
- Minimized risk transfer, with no net risk transfer onto disadvantaged or vulnerable groups.
- Vanishingly small instances of negligent computer driver behaviour, with an additional burden to perform competently at defensive driving skills.
- Conformance to industry consensus safety standards.
- Absence of unreasonable risk (AUR) and/or ALARP risk mitigation on a behaviour-by-behaviour basis.
- Address ethical and equity issues, including transparency, accountability, and absence of a moral crumple zone strategy.

One could argue that this sets a template for other applications of safety critical automation beyond automotive that involve impinging on a human operator's ability to meet their duty of care to other stakeholders.

References

ANSI/UL 4600. (2023). *Evaluation of Autonomous Products*. ANSI/UL 4600, 3rd Edition, approved by the American National Standards Institute, 2023. Underwriters Laboratories, Chicago. https://www.shopulstandards.com/ProductDetail.aspx?productid=UL4600 Accessed 16th January 2024.

AVSC. (2021). *AVSC Information Report for Adapting a Safety Management System (SMS) for Automated Driving System (ADS) SAE Level 4 and 5 Testing and Evaluation*. Automated Vehicle Safety Consortium (AVSC) AVSC000007202107. https://www.sae.org/standards/content/avsc00007202107/. Accessed 16th January 2024.

Bieber C. (2023). *First Offense DUI: Everything You Need To Know*. Forbes Advisor https://www.forbes.com/advisor/legal/dui/first-offense-dui/. Accessed 16th January 2024.

Bidarian N. (2023). *Regulators give green light to driverless taxis in San Francisco*. CNN.COM, August 11, 2023. https://www.cnn.com/2023/08/11/tech/robotaxi-vote-san-francisco/index.html. Accessed 16th January 2024.

BMVI. (2017). *Ethics Commission Report: Automated and Connected Driving*, Bundesministerium für Verkehr und digitale Infrastruktur, the Federal German Ministry of Transport and Digital Infrastructure. https://perma.cc/6UBX-KH5G. Accessed 6th January 2024.

Burton S., Habli I., Lawton T., McDermid J., Morgan P., and Porter Z. (2020). *Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective*. Artificial Intelligence, Vol. 279, Feb. 2020, 103201. Preview available at: https://doi.org/10.1016/j.artint.2019.103201. Accessed 6th January 2024.

CA DMV. (2022). *Report of Traffic Collision Involving an Autonomous Vehicle, June 10, 2022*. State of California Department of Motor Vehicles. https://www.dmv.ca.gov/portal/file/cruise_060322-pdf. Accessed 6th January 2024.

CDEI. (2022). *Policy Paper: Responsible Innovation in Self-Driving Vehicles*. UK Department for Science, Innovation and Technology Centre for Data Ethics and Innovation. 19 August 2022. https://www.gov.uk/government/publications/responsible-innovation-in-self-driving-vehicles. Accessed 6th January 2024.

Congress. (2023). *Hearing: Self-Driving Vehicle Legislative Framework: Enhancing Safety, Improving Lives and Mobility, and Beating China*. Subcommittee on Innovation, Data, and Commerce, Committee on Energy and Commerce. US House of Representatives, July 26, 2023. https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=116277. Accessed 16th January 2024.

Dazio S., and Krisher T. (2023). *As a criminal case against a Tesla driver wraps up, legal and ethical questions on Autopilot endure*. Associated Press, August 15, 2023. https://apnews.com/article/tesla-autopilot-los-angeles-d02769ba359cf6381dc1176c3f5a72a5. Accessed 16th January 2024.

Elish M. C. (2019). *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*. Engaging Science, Technology, and Society (ISSN 2413-8053), Volume 5, pp 40–60 https://estsjournal.org/index.php/ests/article/view/260/177. Accessed 16th January 2024.

European Commission. (2020). *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility*. European Commission Directorate-General for Research and Innovation. https://data.europa.eu/doi/10.2777/035239. Accessed 6th January 2024.

European Commission. (n.d.). *Frequently Asked Questions — Type Approval of Vehicles.* European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs https://single-market-economy.ec.europa.eu/sectors/automotive-industry/technical-harmonisation/faq-type-approval-vehicles_en. Accessed 6th January 2024.

Fleischer M. (2023). *Watch S.F. traffic officers try to get this stuck autonomous Cruise car to move*. San Francisco Chronicle, August 3, 2023. https://www.sfchronicle.com/opinion/article/san-francisco-police-self-driving-cars-cruise-18277009.php. Accessed 16th January 2024.

Gitlin J. (2022). *Tesla recalls 53,822 cars because they won't stop at stop signs*. Ars Technica, February 1, 2022. https://arstechnica.com/cars/2022/02/tesla-recalls-53822-cars-because-they-wont-stop-at-stop-signs/. Accessed 16th January 2024.

Goodall N. (2021). *Potential Crash Rate Benchmarks for Automated Vehicles*. Transportation Research Record, 2675(10), pp. 31–40. https://doi.org/10.1177/03611981211009878. Accessed 16th January 2024.

Goodall N. (2023). *Normalizing crash risk of partially automated vehicles under sparse data*. Journal of Transportation Safety & Security, 16:1, pp. 1–17. https://www.tandfonline.com/doi/full/10.1080/19439962.2023.2178566. Accessed 16th January 2024.

Hawkins A. (2022). *The federal government's Tesla Autopilot investigation is moving into a new phase.* TheVerge,com, June 9, 2022. https://www.theverge.com/2022/6/9/23161365/tesla-autopilot-nhtsa-crash-investigation-emergency-vehicle. Accessed 16th January 2024.

Hawkins A. (2023). *Robotaxis are driving on thin ice*. TheVerge,com, August 15, 2023. https://www.theverge.com/2023/8/15/23831170/robotaxi-cpuc-sf-waymo-cruise-traffic-halt. Accessed 16th January 2024.

HSE. (n.d.). *ALARP "at a glance"*. UK Health and Safety Executive. https://www.hse.gov.uk/enforce/expert/alarpglance.htm. Accessed 16th January 2024.

IEEE 7000. (2021). *IEEE Standard Model Process for Addressing Ethical Concerns during System Design* (September 15, 2021). Institute of Electrical and Electronics Engineers, Piscataway, NJ.

ISO 21448. (2022). *Road vehicles — Safety of the intended functionality*. ISO 21448:2022. International Organization for Standardization, Geneva. https://www.iso.org/standard/77490.html. Accessed 16th January 2024.

ISO 26262. (2018). *Road vehicles — Functional safety*. ISO 26262, in 12 parts, 2nd Edition, 2018. International Organization for Standardization, Geneva. https://www.iso.org/standard/68383.html. Accessed 16th January 2024.

ISO/SAE 21434. (2021). *Road vehicles — Cybersecurity engineering*. ISO/SAE 21434:2021. SAE International, Pittsburgh and International Organization for Standardization, Geneva. https://www.iso.org/standard/70918.html. Accessed 16th January 2024.

Jonsson A., and Voorneveld M. (2018). *The Limit of Discounted Utilitarianism*. Theoretical Economics 13, pp. 19–37. https://onlinelibrary.wiley.com/doi/pdf/10.3982/TE1836. Accessed 16th January 2024.

Kalra N., and Groves D. (2017). *The Enemy of Good: Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles*. RAND Corporation Research Report RR-2150-RC. https://www.rand.org/pubs/research_reports/RR2150.html. Accessed 6th January 2024.

Koopman P., Kuipers B., Widen W., and Wolf M. (2021). *Ethics, Safety, and Autonomous Vehicles*. IEEE Computer, December 2021, pp. 28–37. https://ieeexplore.ieee.org/document/9622307. Accessed 16th January 2024.

Koopman P. (2022). *How Safe Is Safe Enough?: Measuring and Predicting Autonomous Vehicle Safety*. Independently Published, September 2022. ISBN: 979-8848273397.

Krisher T. (2023). *US probes crash involving Tesla that hit student leaving bus*. Associated Press, April 7, 2023. https://apnews.com/article/tesla-school-bus-student-hurt-firetruck-d282a5dd63874f22f5e1a6fc8168801b. Accessed 16th January 2024.

Law Commission. (2022). *Automated Vehicles: Joint Report*. Law Commission of England and Wales, and Scottish Law Commission, HC 1068 SG/2022/15. https://www.scotlawcom.gov.uk/files/4616/4313/7041/Automated_vehicles_joint_report_cvr_24-01-22.pdf. Accessed 6th January 2024.

LII. (2023). Negligence. Definition from Legal Information Institute, Cornell Law School. https://www.law.cornell.edu/wex/negligence. Accessed 6th January 2024.

Liu P., Yang R., and Xu Z. (2019). *How Safe Is Safe Enough for Self-Driving Vehicles?* Risk Analysis, Vol. 39 No. 2, pp. 315–325. https://doi.org/10.1111/risa.13116. Accessed 6th January 2024.

Musumeci N., and Guenot M. (2023). *OceanGate wrote that certifying its sub would block innovation. A longtime sub expert says the 'exact opposite is true'*. Insider, July 25, 2023. https://www.insider.com/oceangate-ceo-certification-innovation-sub-expert-opposite-is-true-2023-7. Accessed 16th January 2024.

NHTSA. (2022a). *Untitled*. Part 573 Safety Recall Report 22V-037, January 27, 2022. US DOT National Highway Traffic Safety Administration. https://static.nhtsa.gov/odi/rcl/2022/RCLRPT-22V037-4462.PDF. Accessed 16th January 2024.

NHTSA. (2022b). *Occupant Protection for Vehicles With Automated Driving Systems: Final Rule*. US DOT National Highway Traffic Safety Administration, March 30, 2022. https://www.federalregister.gov/documents/2022/03/30/2022-05426/occupant-protection-for-vehicles-with-automated-driving-systems. Accessed 16th January 2024.

NHTSA. (2023a). *Early Estimate of Motor Vehicle Traffic Fatalities in 2022*. US DOT National Highway Traffic Safety Administration Report No. DOT HS 813 428. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813428. Accessed 16th January 2024.

NHTSA. (2023b). *Federal Motor Vehicle Safety Standards and Regulations*. (Title 49 Code of Federal Regulations) US DOT National Highway Traffic Safety Administration. https://icsw.nhtsa.gov/cars/rules/import/FMVSS. Accessed 16th January 2024.

NHTSA. (2023c). *Untitled*. Part 573 Safety Recall Report 23V-085, April 11, 2023. US DOT National Highway Traffic Safety Administration. https://static.nhtsa.gov/odi/rcl/2023/RCLRPT-23V085-9893.PDF. Accessed 16th January 2024.

NHTSA. (2023d). *Standing General Order on Crash Reporting — For incidents involving ADS and Level 2 ADAS*. Amended April 2023. US DOT National Highway Traffic Safety Administration. https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting. Accessed 16th January 2024.

NHTSA. (n.d.). *Understanding NHTSA's Regulatory Tools: Instructions, Practical Guidance, and Assistance for Entities Seeking to Employ NHTSA's Regulatory Tools*. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/understanding_nhtsas_current_regulatory_tools-tag.pdf. Accessed 16th January 2024.

Nicholson J., Luttropp D., Jones N., and Friedlander J. (2023). CPUC Status Conference: Safety Issues Regarding Driverless AV Interactions with First Responders. (slides presented at California Public Utilities Commission meeting, August 7, 2023). https://www.sfmta.com/sites/default/files/reports-and-documents/2023/08/2023.08.07_cpuc_status_conference_8.7.2023_final.pdf. Accessed 16th January 2024.

NTSB. (2019). *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018.* National Transportation Safety Board Highway Accident Report NTSB/HAR19/03, November 19, 2019. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf. Accessed 6th January 2024.

NTSB. (2022). *Automobile Safety Rating System Is Failing Consumers*. National Transportation Safety Board, June 3, 2022. https://www.ntsb.gov/news/press-releases/Pages/NR20220603.aspx. Accessed 16th January 2024.

PennDOT. (2022). *2021 Pennsylvania Crash Facts & Statistics*. Bureau of Maintenance and Operations of the Pennsylvania Department of Transportation https://www.penndot.pa.gov/TravelInPA/Safety/Documents/2021_CFB_linked.pdf. Accessed 16th January 2024.

Roy A., Levine D., and Jin H. (2023). *Tesla wins bellwether trial over Autopilot car crash*. Reuters, April 22, 2023. https://www.reuters.com/legal/us-jury-set-decide-test-case-tesla-autopilot-crash-2023-04-21/. Accessed 16th January 2024.

SAE J3018_202012. (2020). Safety-Relevant Guidance for On-Road Testing of Prototype Automated Driving System (ADS)-Operated Vehicles. J3018, 3rd Edition, 2020. SAE International, Pittsburgh. https://www.sae.org/standards/content/j3018_202012. Accessed 16th January 2024.

SASWG. (2022). *Safety Assurance Objectives for Autonomous Systems*. Safety Critical Systems Club — Safety of Autonomous Systems Working Group, January 2022. https://scsc.uk/r153B:1?t=1. Accessed 6th January 2024.

Smiley L. (2023). *The Legal Saga of Uber's Fatal Self-Driving Car Crash Is Over*. Wired.com (July 28, 2023). Available at: https://www.wired.com/story/ubers-fatal-self-driving-car-crash-saga-over-operator-avoids-prison/. Accessed 16th January 2024.

Stewart J. (2018). *Why People Keep Rear-Ending Self-Driving Cars*. Wired, Oct. 18, 2018. https://www.wired.com/story/self-driving-car-crashes-rear-endings-why-charts-statistics. Accessed 16th January 2024.

Stumpf R. (2021*). Feds Order Tesla to Justify OTA Autopilot Updates Instead of Recalling Cars*. TheDrive.com, October 14, 2021 https://www.thedrive.com/tech/42736/feds-order-tesla-to-justify-ota-updates-instead-of-recalling-cars. Accessed 16th January 2024.

UK DfT. (2022). *National Statistics: Reported road casualties Great Britain annual report: 2021*. UK Department for Transport. https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2021/reported-road-casualties-great-britain-annual-report-2021. Accessed 16th January 2024.

United Nations. (2021). *Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems*. UN Regulation No. 157, ECE/TRANS/WP.29/2020/81. https://unece.org/sites/default/files/2021-03/R157e.pdf. Accessed 16th January 2024.

US CFR. (1992). *Reporting of unreasonable risk of serious injury or death*. U.S. Code of Federal Regulations, 16 CFR § 1115.6. https://www.law.cornell.edu/cfr/text/16/1115.6. Accessed 16th January 2024.

Victor T., Kusano K., Gode T., Chen R., and Schwall M. (2023). *Safety Performance of the Waymo Rider-Only Automated Driving System at One Million Miles*. Waymo LLC. https://storage.googleapis.com/waymo-uploads/files/documents/safety/Safety%20Performance%20of%20Waymo%20RO%20at%201M%20miles.pdf. Accessed 16th January 2024.

Wansley M. (2022). *Regulating Driving Automation Safety*. 73 Emory Law Journal (forthcoming 2024), Cardozo Legal Studies Research Paper No. 689, August 15, 2022. https://ssrn.com/abstract=4190688. Accessed 16th January 2024.

Widen W. H. (2022). *Highly Automated Vehicles & Discrimination Against Low-Income Persons*. North Carolina Journal of Law and Technology, Vol. 24, No. 1, University of Miami Legal Studies Research Paper No. 4016783. https://dx.doi.org/10.2139/ssrn.4016783. Accessed 16th January 2024.

Widen W. H. (2023). *Automated Vehicles, Moral Hazards & the 'AV Problem'*. 5 Notre Dame J. Emerging Tech.1 (2023), University of Miami Legal Studies Research Paper No. 3902217. https://dx.doi.org/10.2139/ssrn.3902217. Accessed 16th January 2024.

Widen W. H., and Koopman P. (2022). *Autonomous Vehicle Regulation and Trust*. UCLA Journal of Law & Technology, Spring 2022, Volume 27, No. 3. http://dx.doi.org/10.2139/ssrn.3969214. Accessed 16th January 2024.

Widen W. H., and Koopman P. (2023a). *Winning the Imitation Game: Setting Safety Expectations for Automated Vehicles*. 25 Minn. J. L., Sci. & Tech. 113. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4429695. Accessed 22nd January 2024.

Widen W. H., and Koopman P. (2023b). *Level 3 Automated Vehicles and Criminal Law*. JURIST — Academic Commentary. https://www.jurist.org/commentary/2023/08/widen-koopman-automated-vehicles-criminal-law. Accessed 16th January 2024.

This collation page left blank intentionally.

ISSN 2754-1118 (Online)
ISSN 2753-6599 (Print)

Principles of Conceptual Analysis for Electrotechnical Terminology (ConcAn)

Peter Bernard Ladkin

Causalis Ingenieurgesellschaft mbH, Bielefeld, Germany

Abstract

The term Conceptual Analysis has a storied history in philosophy, in particular in the 20th Century. I argue here for a specific application to electrotechnical terminology, ConcAn. I suggest the necessity for ConcAn by considering a contemporary example from the IEC. I then elucidate some principles of, and illustrate, ConcAn by means of a running example, the notion of reliability. *There are likely important principles still to discover.*

1 Introduction

1.1 Background

In October 2023, a proposal was circulated by the IEC[1] to revise the entries in the International Electrotechnical Vocabulary (IEV n. d.) for the terms "electric" and "electrical". The key entries are 151-11-03 electric: "*containing, producing, arising from, or actuated by electricity*", and 151-11-05 electrical: "*pertaining to electricity, but not having its properties or characteristics*" [2].

I had not thought about these terms before. They are surely basic adjectives in the engineering science of electrotechnology. That suggests to me that we ought to have them clearly and accurately defined for such use.

A non-native-English-speaking colleague suggested that he was not at all clear on any difference. If that were to be so, surely, we could dispense with one of them? Indeed, IEC rules say we should — synonyms are not to be assigned separate entries in the IEV, but just one entry with the synonyms listed there as such.

A look in the Oxford English Dictionary[3] reveals that they register 175 meanings for the English word *electric* and *"[t]he usual current sense*" (no pun intended) is *"[p]owered by electricity, operating by means of electricity; (also) used for measuring or producing electricity*". It also hints that it can be a synonym for "*electrica*l". The OED registers — attempts to register — all uses of the term in the language. Terms used in science and engineering may well have more circumscribed meanings that are appropriate for science and engineering use. We shall see that this may well be so for "*electric*".

[1] The International Electrotechnical Commission, an international standards organization.

[2] The International Electrotechnical Vocabulary is available at www.electropedia.org and entries may be found by searching for the term, or by giving the reference number.

[3] Available at www.oed.com

The OED "*usual current sense*" does not include "*actuated by electricity*", which the IEV definition includes. Let us consider "*actuated by electricity*". My building heating is usually described as "*gas heating*" and it includes a "*gas boiler*", which heats water by burning gas piped in from the street gas supply (heat exchange), and this water is circulated in the building radiators. It is actuated by electrical equipment and thus by electricity: if there is no electricity, my gas boiler does not work. According to the IEV definition, my gas heating is electric.

The IEV definition does not include "*operating by means of electricity*". But that is how my brain works (according to neurophysiologists); indeed, how my nervous system works. My nervous system is "*actuated by electricity*". When Walt Whitman wrote "*I Sing the Body Electric*", he was not writing an ode to neurophysiology[4]; he was trying to say something new.

Furthermore, it turns out all the cars we now have on the roads are really *electric cars* after all, for they are *actuated by electricity*, and *operated by means of electricity*, as anyone knows who has had the misfortune to deal with a faulty generator on a journey, or who has had to call the emergency road service to help with a dead battery. All those diesel locomotives we have on our railroads worldwide are also *electric*, for the traction is provided by motors which are *powered by electricity*. (In GB, they are indeed called diesel-electric, but this moniker is not common in other European countries, or in the US.)

If we take these definitions/explanations of the term as literally meaning what they say, we can thus quickly get into confusion concerning their appropriate use in English. Current usage says: my body is not electric; its source of power is predominantly electromagnetic (EM) (heat ultimately from the sun) and biochemical. My gas heating is not electric; it is gas (that is why we call it "gas heating"). Those internal combustion engine (ICE) cars are not electric, their source of power is chemical, from liquid petrol or diesel fuel. Those diesel locomotives burn diesel fuel as a chemical source of energy. All of this technical material about power and sources and conversions is readily apparent to, say, a bright Chinese-native-speaking engineer. If that engineer is trying to learn English electrotechnical vocabulary, in order to converse with other engineers in other countries, she will be misled by the IEV and OED definitions.

I think we can take it that it is no purpose, either of the IEV or of the OED, to sow confusion about the appropriate use of terms in the English language. It is a purpose of the OED to register all usage; it should not surprise us that some meanings are inconsistent with others. However, for the IEV, we can surely see a purpose for precision and clarity in the meanings given for terms which it takes to be technical. But we have just seen how confusion can easily happen.

What is the remedy? The remedy is, surely, to get it right.

How do we get it right? The proposal of this paper is: conceptual analysis.

Yes, bodily functions such as thinking and moving are actuated by electricity, but the body is not in common parlance electric. Why not? Well, to get our energy, we eat and drink, and we wear clothes to keep warm (37° C is recommended). Eating and drinking provides energy (both warmth and electrical energy) through biochemical processes; the sun and other EM sources provide warmth (and clothing helps us to preserve it). My gas heating is not in common parlance electric. Why not? Because most of the energy which goes to warm my building comes from burning gas and heat exchanging (first to water circulation, then to aerial convection and conduction and EM heat radiation). Similarly, when our IEC

[4] As one might clearly read at https://www.poetryfoundation.org/poems/45472/i-sing-the-body-electric

cars run out of energy, we fill them with burnable, i.e. chemically transformable, liquid petroleum or diesel fuel. When our diesel locomotive needs to run another 500km, it is filled with burnable liquid diesel fuel.

In all of these cases, energy is transformed amongst different types. I would suggest: we tend to label the operation with the *source* of its used energy. "Source" here means "from outside the entity itself". The electricity in the human body, in the diesel locomotive, in my building's heating, in the ICE car, is an intermediary store at best. The source is elsewhere. In the case of my body, there are two: basic environmental heat (we may take it: radiative and conductive) and biochemical. In the case of my heating and ICE and diesel vehicles, chemical.

1.2 Being More Precise About "Electric"

Let me propose: something is *electric* if electricity is the *main source of power for its function*. (In such a definition, *electric* is the definiendum; *main source of power for its function* is the definiens.)[5]

The terms in this definition which are not purely syntactic are *main*, *source*, *power*, *function*. The terms which are "purely syntactic" are *of*, *for*, *its*; their semantic roles are to be explained in terms of how they are used to combine the other terms. Let me call these conjunctive words. We have addressed the roles of conjunctive words in definitions using a technique we called "SemAn" (Ladkin et al. 2023). What I am here calling conceptual analysis considers what the other terms *main*, *source*, *power*, *function* in the definiens — let me call them *conceptual terms* — contribute to the meaning of the term *electric*.

I deal now with the conceptual terms in turn, but in the order *power*, *source*, *main*, *function*.

The main substantive in the definiens is *power*. Such power could be electric, chemical, physico-chemical, e.g. internal-combustive, biochemical, nuclear, laser, heat, light, magnetic. It is surely, and clearly, important engineering-scientific information that the power enabling the function of an entity is, for example, electric.

The definiens says, not that electricity is the *power*, but that it is the *source of power*. Why this? I will take the answer to have been given in the discussion above. Distinguishing between electricity as the *power* and electricity as the *source of power* is thus important to conform with current technical usage.

The term *main* in *main source* is important. For, consider the case of the human body. I take it that it is possible to argue that the source of power for most human functions is electric — the operation of neurons and the complicated nervous-system interactions that give us our human characteristics are predominantly electrical. But the amount of electrical energy involved is small. The main source of bodily power is chemical (food) and heat, as noted above.

Can we speak of everything as having a function? A word for this is teleological (which I have used in systems engineering for nearly thirty years; it hasn't caught on). All engineering-devised systems are teleological; someone put them together for a purpose. The classic von Bertalanffy systems are not teleological — predator-prey systems for example (unless one adheres to a religion which says they are divinely purposed).

[5] Note that this definiendum/definiens pair does not satisfy the substitutability condition: that the definiens be substitutable for the definiendum *salva veritate* in all contexts. A suitable definiens which better satisfies the substitutability condition would be "using electricity as the main source of power for its function".

The term function is also used informally when explaining how a living creature or parts of it fits into its environment; e.g., what is the function of the eye? A strict evolutionary theorist would say that no evolved living creature or its parts has literally a function in this sense; it merely has parts which have had survival advantages over others. But, we can ask, what are these advantages? And, if we like, understanding that proposed mechanism, we can call them by analogy functions, which is what indeed many biologists do.

Consider an art installation that relies upon neon light tubes to provide bright colours. It is art; it is also static; however, it does have behaviour (behaviour is a change of state; it can go from off to on and from on to off). Does it have a function? One may say that it was conceived and built by the artist for some reason — say, as with much art, to induce the viewer to engage. Then it is teleological. Is electricity part of that function? Yes; if it is turned on at some point and that is part of what the artist intended. If it is never turned on and the artist did not intend for it ever to be turned on, the answer could well be no. According to the proposed definition, it is an electric art installation in the first case; not in the second.

I am concerned here, as are most readers of this journal I take it, with electrotechnology — electrical engineering — and its vocabulary. I would propose that all engineering is teleological, and it makes sense for any engineering artifact to inquire what its function is (or is intended to be).

1.3 Conceptual Analysis and ConcAn

The reader may or may not agree with all the considerations adduced above. The important aspects for this article are:

- We distinguish in a terminological definition between definiendum and definiens
- The definiens includes conceptual terms and conjunctive words
- The role of conjunctive words has been (partially) addressed by SemAn (Ladkin et al. 2023)
- The conceptual terms may be considered both individually and in conjunction to determine, along with SemAn, what the definiendum means in electrotechnological use
- The analytical consideration of the conceptual terms is what I am calling conceptual analysis
- The specific form of conceptual analysis advocated in this paper is called ConcAn

There are some things to be said about electrical*, and how this term differs from* electric*, but I would suggest that is for terminologists improving the IEV and not for this paper. In this paper I propose a way to go about analysing concepts, as I did electric, above, and thereby improving the terminological definitions available to us in electrotechnology. I call this* "ConcAn".

2 Some Philosophical History

2.1 Conceptual Analysis

The term conceptual analysis has a history in philosophy, which can be argued to originate with Immanuel Kant (although the techniques are as old as philosophy itself). This section gives a brief survey. The point of such a survey is that the history of conceptual analysis

in philosophy includes major contributions by major intellects, and the reasons why those contributions arose is important to understand what they offer. Much of it will not necessarily be relevant for electrotechnology, but some of it I would argue is not only relevant but very important.

2.2 A Very Brief Summary of Conceptual Analysis in Analytical Philosophy

It is fairly clear to most people what a physical object is. They are encountered every minute of our waking life, and indeed our sleeping life also, although we are less aware of them. We can designate them and communicate about them when they are not ostensively present. But those same communicative actions can also be used to communicate about things which may not exist in the same way, or at all. I can talk about a ghost the same way I can talk about my next-door neighbour. I can identify patterns of objects: there is an orange and another orange; there is an apple and another apple; there is something similar, we think, about those two situations. We call that similarity the number of designated objects, in this case two. But I cannot point to a number itself, such as *two* or *three*, or throw it to you, or eat it, in the same way in which I can an orange; the mode of designation, we might call it, is different.

British "empiricists" in the 17th and into the 18th Centuries (well exemplified by John Locke, Bishop George Berkeley, and David Hume) were concerned to explain "the world", i.e. the world of physical objects plus whatever, through considering how our mental activity connected to the world of objects independently of our bodily capabilities. Features — patterns — in the world of objects were "perceived" by us. According to Locke, there were supposed to be mental correlates of these features somehow "in the mind" of the perceiver (when perceiving veridically); these mental correlates were called "ideas". We have immediate cognitive access to the ideas occurring in our minds; according to Locke's philosophical psychology that is *exactly* what we had cognitive access to. But then the problem became to explain veridicality: if our ideas are all we have access to, how could we determine they were veridical representations of the world? For we had, by hypothesis, no access to "the world" in itself to compare, just to our ideas. This came to be called the "veil of perception", behind which we could, by hypothesis, not peek.

Bishop Berkeley bit the bullet and said, actually, "the world" consists of ideas in the mind of God, and the way we know our ideas (perceptions) are veridical is because the good God ensures that it is so in various ways; he doesn't try to deceive us, nor does he let us be deceived about them, as Descartes worried. Berkeley's is an answer of a sort, but just pushes the problem a bit further: how do we know that God is like that? One answer, "faith", is Christian-religiously but no longer philosophically respectable — the question proved to be almost impossible to answer for those not disposed to Christian-theological precepts. Berkeley's position came to be seen as a version of "assume the answer".

David Hume worried (amongst other things) about how "the world" exhibited its regularities. How and why do things happen? What are the origins of such happenings — the causes, as they came to be called — of things which happen? We can see causes and their effects as "constant conjunction": when phenomenon A causes phenomenon B, A is always temporally accompanied by or followed by B. Hume noted that we can observe constant conjunction, but what we cannot do is perceive causation happening directly. But if we cannot observe causation directly, how can we be sure that the next iteration of A will be followed by B? We can, if it is indeed the case that A causes B, but we cannot be sure of that from observation — we only observed the constant conjunction; maybe what we observed was just happenstance, and not an instance of causation? The sun has risen

each and every day of the world's existence; elaborate theories of the causes of that phenomenon, from Ptolemy through Copernicus and Kepler to Newton, were/are available. But what if tomorrow it didn't happen? How do I know that I have been seeing an instance of causation? How can I reason from what has happened to what will happen?

Kant attempted to shoehorn all this into structural constraints on perception and cognition, and to derive from them, along with veridical perception, concrete manifestations of causality. These structural constraints could be called concepts, and thinking according to these constraints is conceptualisation. These concepts were to be independent of people, in so far as they were still to be present even if there were no people around to engage in perception. One might say, the structure is valid even if not realised. Our perceptions are, and experience is, veridical in so far as it conforms with these constraints. When it does, causality is manifested as such-and-such. It is facile, one might think, to summarise Kant's philosophy in a few such sentences as these. I would agree, but Kant's philosophy is not the subject of this paper. Rather, his notion of concept and the kinds of analysis in which he indulged can be considered the modern origins of the topic of this paper.

Kant considered in detail what it meant for a *concept* to have a definition. For example, one could define a bachelor to be an unmarried man. How does such a definition work? Kant suggested such assertions are *analytic*, meaning for him that the "concept" of bachelor "contains" the "concept" of "unmarried" as well as the concept of "man". The notion of "contains" needs elaboration (and will not get it in this document). A statement/assertion for Kant is a statement P(A), where A denotes the subject and P the predicate. The idea would be that the predicate "contains" the concept under which the subject is presented. So this means that, in P(A), which is analytic, A is not a proper name, but a presentation of a kind of thing: "a bachelor", for example, which is a count noun; or a mass term — "water", for example. Thus, "Peter is an unmarried man" could not be analytic because of its form ("Peter" is a name, not a count noun or mass term); "a bachelor is an unmarried man" is analytic[6].

Kant claimed all assertions had such subject-predicate form. This view changed with the introduction of the language of first-order logic (FOL) by Frege, and its subsequent uses, for example in Russell and Whitehead's *Principia Mathematica*, and Russell's *Theory of Descriptions*, in which an assertion can have the general forms known in FOL[7]. So, for example, "all bachelors are unmarried men" could be analytic, even though it is not nowadays seen to have subject-predicate form; c.f. "all positive integers are either even or odd". Conceptual analysis thus could come to mean the study of analytic statements and what makes them analytic, and how to tell if a statement is analytic.

In the mid-20th Century, W. V. O. Quine wrote an influential paper, *Two Dogmas of Empiricism*, in which he cast doubt that the Kantian notion of analyticity had a well-defined meaning (Quine 1951). This is not quite the way he put it, for he cast doubt on the coherence of the notion of unambiguous "meanings" in general. In his later book, *Word and Object*, he illustrates in detail through the notion of "radical translation" (Quine 1960). He considers in detail an attempt by an individual to learn an unknown language from scratch. Watching a speaker use so-called "observation sentences" is helpful: observation sentences are assertions about things in the present environment to which one can point with a hand or a finger. A speaker points at a rabbit and says "*gavagai*". Quine argues

[6] There is a question whether "a bachelor is a married man" should be considered analytic in Kant's sense, for the subject is not "contained in" the predicate, but rather explicitly excluded by it. Modern discussions which accept the notion of "analytic" would deem the sentence to be *analytically false*; in this case the relevant predicates of sentences are "analytically true" and "analytically false" and the term "analytic" can be taken to mean "either … or …" .

[7] This view has since widened further to include more complex formal-logical languages than FOL.

that a "radical translator" is fundamentally unable to distinguish whether "gavagai" is a rabbit, and the assertions about "gavagai" assertions about rabbits, or whether "gavagai" refers rather to a coherent collection of mutually cooperating rabbit-parts, and assertions about gavagai are assertions about mutually-cooperating rabbit-parts. He further observes that these two kinds of discourse in our language are logically different: one assertion is about a singular (rabbit); the other about a plurality (lots of rabbit parts). This is the thesis of the "indeterminacy of translation": gavagai is a rabbit, or equivalently many cohering rabbit-parts, and according to Quine a translator fundamentally has no means of telling which ontology is being used.

It follows that this must be so even when you and I are talking to each other in the "same" language: you cannot fix my "set of concepts" just by experiencing how I use them, and there is no way of eliminating this indeterminacy. It follows further that, if there is no way of telling what concepts are being used and how, there is no reasonable way in which we can talk to each other about concepts, except in so far as we make "working assumptions": when you say "*rabbit*", I presume your concept is identical to my concept of rabbit. That makes it easier to work with you, but it by no means entails that you have my concept of rabbit. So that sets the whole idea of "concepts" in question. Quine held that this even extends to logic and mathematics: he further suggests that, if even these are indeterminate, we might just as well use first-order logic (FOL) and mathematics based on it (and Peano Arithmetic, and…).

If there are no concepts as construed in broadly Kantian terms, then there is clearly no conceptual analysis based on broadly Kantian notions. The "standard story" (one version of it) is then that Saul Kripke got the notion of "concept" back on track. He became known in the mid-1950's for a piece of logic performed when he was a teenager. He gave a Tarskian semantics for so-called modal logics; formal logics that, besides the sentential logical operators AND, OR, NOT and IF...THEN, and quantifiers if first-order, have additional unary sentential operators NECESSARILY and POSSIBLY (which are "dual" in a technical sense: an assertion is necessary if and only if its negation is not possible, and something is possible if and only if its negation is not necessary). Quine had rubbished modal logic, by connecting it with the notion of "analytic", which notion he had argued could not be coherently sustained. Kripke gave a "possible world" semantics, a Tarskian model-theoretic semantics in which the models of a theory in modal logic were a class of "possible worlds". Some of those worlds were "accessible" from others, and some not. Those sentences (NECESSARILY A) were true in a world W just in case A is true in every world accessible from W.

So, for example, a world W.1 in which I am driving a car, swerve to avoid an obstacle and narrowly miss crashing into a wall at 60 kph is accessible from a world W.2 in which I am driving that car, and the same things happen, except that I hit the wall head-on (and all which follows from that) at time T. The world W.3 in which I am alive at T+100 seconds is accessible from W.1 (indeed, W.1 itself might be such a world), but it is implausible to claim that it is accessible from W.2; as we say, "*no one could survive that*", or as we might say using the terminology of modal logic, "*it is not possible to survive that event*".

Kripke extended his analysis in an article (and then a book), *Naming and Necessity*, in which he explained how names came to be used for objects, and how from these processes it could be argued that some objects (referred to by using their names) had some properties necessarily (that is, had those properties in all possible worlds), and these properties could be then taken to define the object that the term named (Kripke 1980). Such terms were called "rigid designators". Concepts and their definitions, according to this telling of the story, were rehabilitated through this account.

This story is well-summarised by Jerry Fodor (2004). Some more matter on Conceptual Analysis in philosophy may be found in Beaney (2003) and in Margolis & Laurence (2005).

2.3 The Importance of This History for Electrotechnology

2.3.1 Preamble

The importance of this story for electrotechnological conceptual analysis is threefold. First, *pace* Quine, conceptual analysis can be performed (Kripke). Second, there are people working in electrotechnology who think that truth is relative (to one's interests — see below for a fairy tale). That truth is relative to one's interests, rather than a measure of an objective condition of the world, is a thesis associated with philosophical pragmatism, an approach to which Quine is commonly held to conform. That truth is relative is a common accompaniment of the Quinean argument, but more often associated with Richard Rorty, and far more prevalent in the academic humanities, and thereby amongst far more people than there are philosophers, let alone philosophers with the academic inclination to refute it.

2.3.2 Contra Analytics: The Relativity of Truth and Its Use in Business

People who hold that truth is relative set little store by careful definition, because there is, according to them, no such thing. To those of us who believe that conceptual analysis is helpful, even necessary, it can be a hindrance to have people propose what we take to be poor or deviant definitions of technical terms and insist it does not matter. A fairy tale suffices to illustrate[8].

The fairy tale: If truth is relative, then it can be deemed by a company's engineers to be relative to the business model of their company. Say the business model says "*As the solution to issue Y, we do X; we do X particularly well; everybody should do X, and they should engage us to do X for them*." The question for a conceptual analyst would be *inter alia* if X makes sense as a solution to issue Y. To a non-relativist, the answer would be maybe it does, or partially does; or maybe it doesn't. To a relativist company engineer, it suffices that the company advocates X as the solution to Y for X to be the solution to Y. To that engineer, there need be no "yes, but …." to which she has to accede. There need be no discussion of whether X is a satisfactory solution to Y in which she has to participate. She has only to align customers with the company's point of view.

2.3.3 Notions in Cybersecurity

An example of such analytics and/or their lack may be seen in the way in which cybersecurity and functional safety is currently handled in IEC standards. There is an IEC Technical Report (TR, which gives recommendations, but does not contain normative injunctions) that recommends how cybersecurity considerations should be integrated into the development of safety-critical E/E/PE systems. This TR, IEC TR 63069:2019, uses centrally the notion of "security environment", which is defined as "*area of consideration where all relevant security countermeasures are in place and effective*". Setting aside

[8] A reviewer wished for examples of relativity. To elucidate concrete instances would likely contravene a number of my professional commitments (including codes of conduct); hence the fairy tale.

what the first four words may mean, and using instead the term in the definiendum, a reasonable interpretation of this is "a system environment in which all relevant security countermeasures are in place and effective". It was discussed in Ladkin (2020) whether (and how) such a notion can make any sense. Reading books for practitioners by some eminent authors could lead one to the conclusion that such an environment cannot exist in the current state of the art (Spafford et al. 2023).

IEC TR 63069 advocates establishing a "security environment" (in its sense) for the development of safety-critical systems. The TR is in process of becoming a Technical Specification (TS), a document which has normative import. That means that those developing or integrating systems with safety-critical components conformant with the document will be required to establish a "security environment", which is an impossible task if one cannot exist. One can imagine, though, companies promoting their expertise in establishing such "security environments" according to IEC TR 63069:2019 and its successor TS and gaining business by so doing. A conceptual analyst conversant with Spafford, Metcalf & Dykstra (2023), and other literature from cybersecurity experts, may well object to such marketing. The company can claim that an IEC standard says do it, and they do it, so you should buy their expertise. The analyst says no one can do it, in particular the company, so no one should believe the misleading claims. The company has "its truth"; the analyst has "her truth". "So what?", says the truth-relativist (amongst them, the company).

End of fairy tale: The reader may conclude that this situation is satisfactory, or that it is unsatisfactory. If this situation is unsatisfactory, part of the goal of this paper is to demonstrate what can be done about it conceptually-analytically.

2.3.4 An Example of Fruitful Application

It turns out that possible-worlds thinking, of the sort introduced by Kripke as part of his rehabilitation of concepts, is amenable to practicing engineers. I introduced a causal-analysis method, Why-Because Analysis (WBA), starting in 1995, to analyse accidents causally (Ladkin 2001) (Ladkin 2017). WBA is based on the counterfactual notion of causality expounded by David Lewis (1973a) and others (Collins et al. 2004) (Paul and Hall 2013), and counterfactual assertions were given a possible-worlds semantics (also by Lewis (1973b)). Our early experience with WBA showed that engineers were able to reason counterfactually in a more or less uniform manner (their judgements on the counterfactuals which occur when trying to analyse engineered-system causality counterfactually were generally uniform, modulo independent sources of error). Because of this, we could base our fault analysis method, Causal Fault Analysis (CFA), as well as our hazard analysis technique, Ontological Hazard Analysis (OHA), also on this approach (Ladkin 2017).

3 Goals of Conceptual Analysis in Electrotechnology

3.1 System Science

Electrotechnologists (and others) nowadays concern themselves with complex conjunctions of equipment, each of which can be individually quite complex, either with notionally complex behaviour, or with complex conditions under which the desired behaviour is to be guaranteed as far as possible. These conjunctions of equipment are

called systems. There is an emerging branch of technology called systems engineering, which is concerned with how systems are built, how they acquire and maintain the properties desired of them by their creators, what their desired and undesired behaviour may consist of, and so forth. Much of this work is nowadays very conceptual, for example the documentation of moderately complex dependable software includes specifications, as well as arguments providing reasons why the software satisfies its requirements, and also lower-level specifications, because such systems are often designed hierarchically (or "recursively" as some system engineers like to say (Ricque 2019)). Such software is mostly built hierarchically unless it is very simple indeed, and most assured-dependable software is not simple, in the sense that even a few lines of code may have complex behaviour when compiled, linked and loaded onto hardware which executes it. The documentation of how all this fits together is equally complex. And then there are test designs and documentation of the results of tests and assessments of test coverage. Even here, testing is known to be scientifically inadequate to assure properties of software that is required to be highly dependable (Littlewood and Strigini 1993).

Such a complex process as critical-software development, assessment and operation, then, is driven by intermediate items ("documentation" in the phrase of David Parnas[9]) which are primarily conceptual. A collection of concepts demonstrably adequate to this task is key to its accomplishment.

System science is not the only phenomenon in modern electrotechnology, but it is currently one of the least developed. The technical terminology for the physics and mechanics used in electrotechnology is in contrast pretty well established (in some cases for a century or longer). The conceptual analysis required is a matter of going back to visit physics courses to refresh it.

Engineers make statements about systems in discussion with other engineers and others, and it is currently hard to tell how many of these statements connect with the real world of a plethora of engineered and other objects engaging in joint behaviour. So, while systems science is not the whole story, we can maybe use it to illustrate how a useful conceptual apparatus may be built up, and what the adverse consequences can be if it is not.

3.2 The Vocabulary of Systems Science

First, systems have parts. It makes sense to call some of those parts systems in their own right (in this case, subsystems[10]) and there are other parts which may be taken to be unitary[11]. Subsystems may overlap — some unitary part may belong to more than one subsystem. So terms are needed for unitary parts, as well as sets of such parts engaging in more complex coordinated behaviour within the system.

1. *A vocabulary of architectural terms (for systems and parts of systems)*

[9] "Documentation" consists at least of descriptions of an engineered system and its intended behaviour, usually on multiple levels from very general ("functional requirements") through intermediate levels (e.g., "functional architecture/specification") down to implementation (the actual system itself). The "levels" are intended to be related to each other through "refinement" and conversely "abstraction". There is not yet agreement on what levels there are or what they are called; whereas refinement is pretty well defined; abstraction is its relational converse.

[10] Although system engineers adhering to the INCOSE terminology use a different term.

[11] "Atomic" is a word philosophers and logicians often use for this, said to derive from a Greek word for "uncuttable". But in electrotechnology applications, especially with modern chip technology, the physics meaning of "atom" surely has priority.

An observation: Currently, in IEC 61508 on E/E/PE functional safety and in the IEC 62443 series on cybersecurity in IACS[12], there seem to be fourteen different terms used for such parts, with various distinguished additional properties (Ladkin 2019).

Second, these parts interact — the system engages in behaviour. Since we are talking of engineered systems, rather than systems which just exist as they are in nature (biological and ecological systems), we can speak of the purpose(s) of a system — it is teleological. Since a system has a particular condition at a particular moment of time (it has a state) and engages in behaviour (a series of events considered together; an event is a change in state of something), the purpose of a system is described using vocabulary appropriate for these states and events/behaviour. Such descriptions are key parts of a teleological system, then. Such an ontology is given in Ladkin (2001).

2. A vocabulary for descriptions of events, states, and behaviour

Teleological systems come with descriptions of (actual or desired) behaviour, but systems and descriptions are very different kinds of thing. The physical constitution of a teleological system is in itself key to its purpose; in contrast, the physical constitution of a description is not — it is not really relevant to the purpose of a description whether it is printed on yellow paper or white paper, size A4 or size A3, whether stapled together, or loosely bound, or held in a folder, or displayed on a computer screen. Descriptions are important for their denotational and semantic content, not their physical attributes. Standard concepts of the semantics of assertions, such as consistency and completeness, expressibility, and so on, are applicable to descriptions of this type.

Third, we have just observed that, in order to deal with teleological systems, we have two categories of thing: system parts themselves and their interconnections; and descriptions. Descriptions are in language; they say things about the system. Given that we have physical objects and descriptions, we can now ask if there is any relation. Yes, we might say, the systems are teleological — we certainly want there to be a relation: a requirements description is supposed to say what people want the system to do (and not to do). So there need not just be system properties and documentation properties but also system+documentation properties, such as concepts, which say how well a system fits its description, or not. Such concepts as "*reliability*" (= it fits, or it mostly fits to some specified degree). Such as concepts which say how characteristics of a system may breach the boundaries of its description ("*emergent properties*"). A technical word in logic for "types of thing" is: *sort*.

3a. A vocabulary of sorts: at least
- *architectural concepts and*
- *descriptions*

Besides system objects, documentation, system and object properties, properties of system-documentation, there is also an environment in which the system operates, generally taken to be a collection of objects which are not part of the system but with which the system interacts (in some of its behaviours, including unwanted behaviour). There are also usually people — assessors, operators, recipients of benefit from system operation, and so on. People may themselves be either individuals or grouped into organisations with specific purposes for, or expectations of, the system. Such people may be real people, or (some of) their interactions may be substituted by robotic agents (for example, telephone switchboards nowadays often do not have human operators). Systems which have functions or sub-functions executed by real people and human organisations are dubbed

[12] IACS = Industrial Automation and Control Systems.

sociotechnical systems. The "socio-" part of a sociotechnical system is generally considered part of the system, so there may be human or robotic agents both within the system operation proper and outside the system, interacting with it. There needs to be vocabulary to describe the environment-system relations and behaviour and the sociotechnical relations and behaviour.

3b. A vocabulary of sorts: including for
- *agents (human and artificial) and their organisation*
- *environment-system relations and behaviour*
- *sociotechnical relations and behaviour*

3.3 Components of a Conceptual Analysis

There are more than these different sorts involved in electrotechnology. They all need commonly-understood terms, a terminology in other words. As we have noted, systems and their parts are important for what they are and what they do; how they are constituted is often (mostly) crucial to their operation. Specifications and descriptions are important for what they say, it is relatively unimportant whether they are pixels on a screen, byte data in memory, or ink on a piece of paper. There are switch mechanisms, temperature, humidity, machine-code programs, bandwidth, assertions, test designs, and operator-cognitive saturation to talk about. These are all fundamentally different kinds of thing, with radically different properties, and different properties we care about. Assertions have logical consistency (or not); switches are not the kind of thing which can have logical consistency. Cognitive saturation is a property of a person (or not), or maybe a robotic agent; machine-code programs are not the kind of thing which can have (suffer) cognitive saturation. Philosophers talk about "categories" of thing, e.g. Ryle's "category mistakes" (Magidor 2019)[13] or "natural kinds" (Kripke 1980); things whose radical differences are to be treated as objective features of them, and not subjective distinctions invented by human observers. As noted above, logicians talk about "sorts". A language and logic which can accommodate different sorts is known as a sorted logic. We have used the formal language of sorted logic (first-order logic with sorts), "LSL", in SemAn (Ladkin et al. 2023).

Is a logic necessary for terminology construction? I would suggest it is. For example, definitions need to be checked for self-consistency. An object (in terminology, mostly a type of object rather than a specific instance) whose definition is self-contradictory[14] does not exist so there would be no point, and no place, in electrotechnical terminology for that definition.

One might suggest that self-contradictory definitions are silly things that will not be proposed by serious engineers. But one cannot necessarily tell that a definition is self-contradictory from its surface construction. In particular, if the definition uses terminology from a logically-formulated mathematical theory, which many terms in electrotechnology do (since much of it derives from the physics of electricity), it may propose something which is inconsistent with that theory, and for most interesting mathematical theories there exists no algorithm which can be relied upon to show it, if so (this is a corollary of various extensions of Gödel's Incompleteness Theorem, in particular one due to Rafael Robinson concerning Q, his "weak" formal theory of arithmetic (Robinson arithmetic. (2023)).

[13] For a much shorter version, see "category mistake" in Blackburn (1994).

[14] Inconsistent" is another term for "self-contradictory" here.

Also, a chain of definitions might well be inconsistent taken as a whole, and it does not take many such definitions to hide a mutual inconsistency effectively. I have experienced something like this. I wrote a conference paper in 1987 which used axioms for a specific theory, proposed by two well-known theoreticians in AI, and derived consequences from it. The paper was published in a prestigious conference (Ladkin 1987a), and went in to a draft of my thesis. When my thesis advisor, Ralph McKenzie, read it, it occurred to him that something was amiss, because the consequences did not match what he thought should be the case. It took him a while to realise where the problem was. One of the axioms was syntactically wrong. I had argued syntactically from the axioms to consequences that Ralph felt could not be what had been intended. Ralph and the original authors had worked from what they thought the axioms should be saying rather than, as I did, from what they in fact said. Anthony Galton published a paper in a prestigious journal analysing the mistake in my conference paper (Galton 1996). It had been of course corrected in my thesis after Ralph had pointed it out (Ladkin 1987b).

Michael Jackson points out (Jackson 2013), with examples, that many sets of requirements for complex systems are inconsistent; even that one can expect them to be inconsistent, given that they are proposed by different groups of domain specialists who do not cross-check each other's usage.

In summary, consistency is currently a poorly-addressed problem. It can be addressed, definitively, through using automated consistency checkers on terminology formulated in LSL (and other logical means).

4. Use LSL, in particular to enable consistency-checking

My colleague Bob Riemenschneider showed that Bertrand Russell's idea of conceptual analysis included what the "arity" of a predicate is (Riemenschneider 1984). The arity of a predicate (the number and sort/type of its arguments) is explicitly exhibited in LFOL[15], whose current notation Russell helped develop. We can illustrate its importance by considering at length the notion of reliability, which is important for electrotechnology.

Reliability means, informally, in engineering that an engineered artefact does what you want; behaves in the way you wish it to behave. Airlines speak of "dispatch reliability"; it means how often an aircraft type can take off on a scheduled flight (if it does not satisfy the conditions for flight, for example some item on the Master Minimum Equipment List (MMEL) is not working, the aircraft may not take off). That a switch is reliable means that it switches a circuit on, and switches that circuit off, when it is activated so to do. Thus described, these are both on-demand functions: an aircraft dispatch (= take off into flight) is a discrete event, as is activating a switch.

An example of continuous-function reliability is the steering of a car. The activation of the steering wheel in the driver's compartment activates a change in angle of the steered road wheels, and there are constraints on the way it is done, so as to fulfil what the driver expects.

5. Determine the arguments (components), their sorts, of a concept, and thereby its arity

Systems are built from components and parts, and can become very complex (modern military and commercial aircraft are built from hundreds of thousands of parts, or more). Physical parts with a function wear out when, for example, they lose matter during use and thereby change shape. When they wear out, they cease to function in the intended way.

[15] The language of first order logic

And they wear out at different rates. How do you keep a system running, when all of its hundreds of thousands of parts have different wear-out behaviour? There is a technique to keep track of these situations, called FMEA[16] (Carlson 2012), which has been very successful in increasing the "dispatch reliability" (as above) of, for example, complex vehicles and aircraft.

Suppose we are to introduce a technical engineering term "reliability" that is supposed to reflect this high-availability facet of usability. The phenomena described above are important — we want our artefacts to behave in the way we want, and we want to develop techniques such as FMEA to enhance this if we can, so we do need to instantiate this concept somehow. As described above, dispatch reliability is a value property (called a fluent in symbolic AI[17]), which has a value usually measured in per-cent. Here, we already have a formal choice to make: could it be a Boolean property which something either has or does not have? The answer is that, purely formally, it does not matter[18]. So there is one result of conceptual analysis: a particular choice of representation is formally inconsequential. This observation is the result of considering the representation of a concept in a formal language.

Considering further the example of "dispatch reliability", or reliability in the FMEA sense, a system (a system part) can be generally expected to work for a time, and then wear out. We could, if we like, gather data on this. We put a switch on a rig which switches it on and off continually and see how many times this can be done before it mostly no longer works. We can do it for lots of switches, so we can estimate an "expected operating life" of the switch type: we can expect the switch to operate for N demands before it doesn't. So, if we install such a switch, can we expect it to operate for (N-1) demands correctly, and then swap it out for a new one? No, some will quit earlier and some will carry on working for more than N demands. So, next observation: we are in the realm of statistics, and statisticians have a good intellectual handle on all this, with their set of concepts which work. We can use them if we wish. Statisticians tell us, after our tests, that we can expect to the kit to work for N demands to a particular *level of confidence*. For example, if N is the (true) mean, then our confidence should surely be 50%. But we can mostly never get at the true mean: we just have our tests, and can estimate the mean from those tests, along with an expected error (or, again, confidence). Connecting statistical-parameter values with confidence in this way is quite well developed — confidence theory for classical statistical inference stems originally from Jerzy Neyman in the 1930's (Neyman 1937).

Note at this point that we are beginning to diverge from what a philosopher would do. A philosopher would say: what are statistics? How are our inferences justified? Is an "expected value" an objective thing, or a mental construction (or both or neither)? Are we Bayesians, for which the numbers tell us our best guess modified by experience, and are thereby subjective[19]; or are we frequentists, for whom the mathematical parameters of a statistical distribution have an objective presence out there in the world? Or are we Laplacians, for which the propensity of a die to come up six is a property on a par with the

[16] Failure Modes and Effects Analysis

[17] See https://en.wikipedia.org/wiki/Fluent_(artificial_intelligence). Accessed 26th January 2024.

[18] We can trivially turn any value predicate into a Boolean predicate by adding an extra argument for its value: when f(x,y,z) = a, we can reformulate as f'(x,y,z,a) = TRUE, along with a condition that f' is functional in its fourth argument: (for any x,y,z)(f'(x,y,z,a) && f'(x,y,z,b) → a=b). And vice versa; any Boolean predicate with at least one functional argument can be turned into a function with one argument fewer.

[19] Bayesians speak rather of "critical interval" rather than "confidence" or "confidence interval". The meaning of critical interval is arguably a more intuitive notion of confidence than the Neymanian version. The reason for the divergent vocabulary is clarity. Neyman bagged the term "confidence" first. Others with a different idea needed a different word.

colour of the shirt I am wearing, which is grey (and not yellow, and not blue). We don't do that in engineering. We say: statistics, well developed, great, let us use it.

It follows that there are at least a number of different sorts in this definition. We have the kit itself. We have some idea of an intended/expected function. We have a number of demands, or a period of time. We have an expectation. And we have a confidence in that expectation. All this, although we are not yet very far into our analysis. Next point: it is well to keep track of all the sorts which are hanging around.

Notice also that we are talking about things we can see and test. Suppose we were to try to connect this concept of reliability with Laplacian propensity: say, the reliability of a six-throw of this die is its propensity to come up six when thrown. We can throw it any number of times, and register the proportion of sixes which arise. Is that the propensity? Surely not, for I get a different proportion on the very next throw (no matter what the result), so the propensity would have changed without anything concerning the constitution of the die having changed in any non-negligible manner, whereas the propensity is supposed to be a property of the die which does not change if the constitution of the die has not changed in any non-negligible manner. It follows that we cannot identify propensity with observed proportion. Even if it is not directly observable, is there some god who will tell us in our dreams what the propensity really is? If there is one, that does not appear to be the way she chooses to work.

It turns out to be hard to connect propensity with any observable property, which is likely why statisticians and probability theorists themselves have given up on the concept. We have Bayesian subjectivists, who say probability is "*your best guess, modified by experience*", and frequentists, whose calculations follow the Humean idea of constant conjunction: run the tests for long enough and limit theorems tell you numbers converge; the limit number is the parameter of interest. But there are no (coherent) Laplacian propensitivists. We want our concepts and calculations to help us make a difference: to help us make this kit work more often, or better. We need to do stuff and observe the effects. Observable effects may lead us to useful concepts that are not themselves directly observable but, to be useful, their values will be manifest in observed/observable effects. Engineers have no use for Laplacian propensity. This might be another point of divergence from philosophers (although there have been and are philosophers who propose that any meaningful concept must be manifest somehow in observable effect — the so-called logical positivists, for example. But they, as well as Popperian refutationists, had a problem explaining how mathematics is thereby meaningful. As well as a problem determining what constitutes "observable"). It might well be that we can say that the engineering approach to meaningfulness is broadly that of logical positivism: concepts useful to engineering are based as far as possible on (naïvely) observable properties. So, another point of conceptual analysis: what is observable, directly and indirectly? Is everything we define based on direct or indirect observability? Do we have any hidden metaphysical notion such as propensity? (If so, let us get rid of it.) Let me call that notion d/i-observability.

6. Use, as far as possible, d/i-observable terms

To determine the reliability of the kit, then, the behaviour of the kit must be observable in such a way that we can say whether it is performing its intended function. We have just looked at observability. Let us look at some of these other terms more closely. "*Performing*", for example. Could we substitute with "*behaving*"? I take "*behaviour*" to be a logically definable term, as in say Ladkin (2001). We can surely say "*behaving according to its intended function*" and mean the same thing as "*performing its intended function*". Here is another point of conceptual analysis, a sort of Occam's razor. Build

concepts as far as possible on what is already well-defined; "new" concepts (such as "*performing*") shall be defined. This entails that concept definition be iterative, as far as possible.

What about "*intended function*"? Here it gets a little tricky.

3.4 Intended Function: An Example

A recent story — it is a little involved, but please bear with me.

The telephone connection of my housemate does not always work (she has a DSL connection for Internet and a WiFi access point, as well as DECT telephone service, all in the same box). There are access points/junction boxes attached to the outside of buildings, called APLs[20]. The telephone-system cabling up to and including the APL is the property of and responsibility of the network operator (pervasively Deutsche Telekom); the system cabling into the building and to the end-devices (telephones and other devices) is the property of and responsibility of, generally, the building owner.

First, a bit of topology (actually topography, but in telecommunications networks it is usually called topology). See Figure 1. My neighbours have an APL which I can see from my office. They are on Road y (Ry). Their house is, say, House v (Hv). My housemate's house line is connected to an APL which I can see to the right if I lean out of my office window on the first floor. She is on Connector 7 of that box. That building, House z, and thus the street address of the APL, is also on Ry.

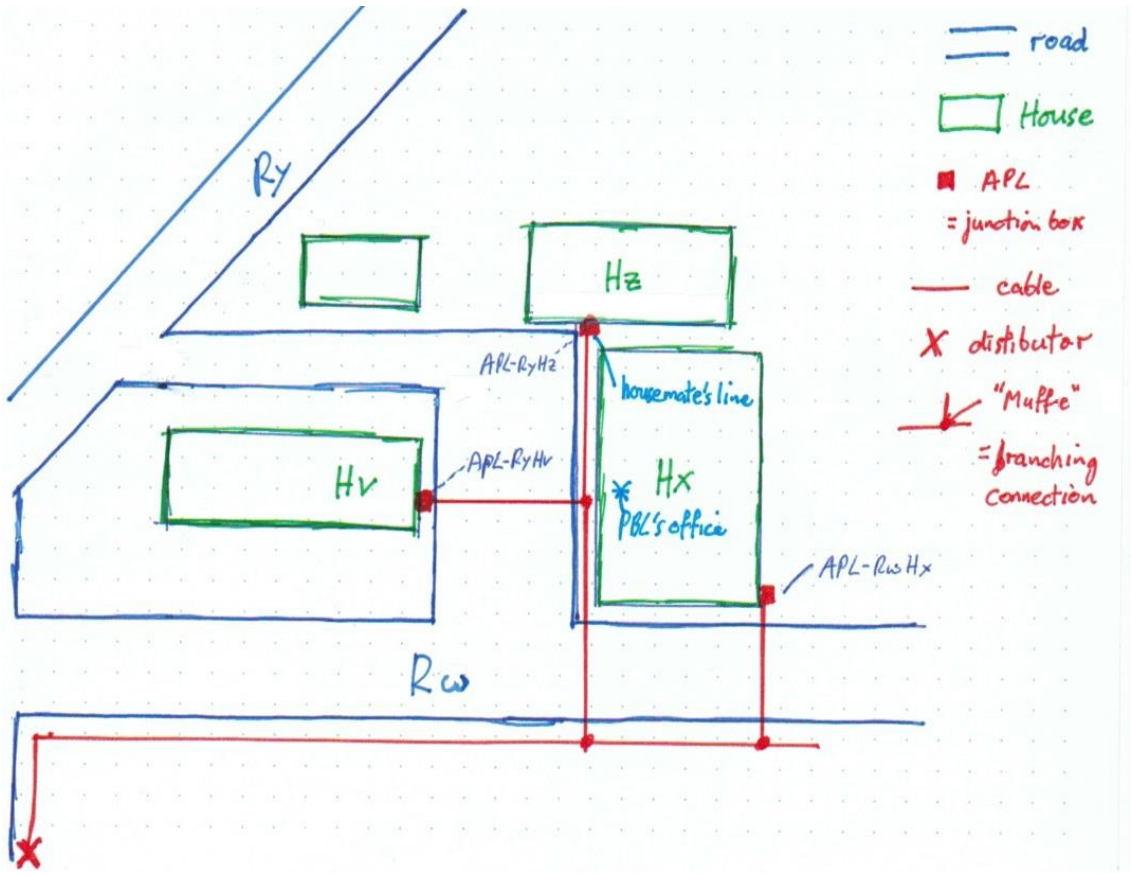

Figure 1 ~ The Street Telecommunications Connection Topology

[20] APL = Abschlusspunkt Linientechnik

I am not on Ry. My house has an address on a different road, Rw. My APL is on the other side of my house from my office, and my two lines, both on Connector 1, run in through the cellar to a point on the ground floor in the SE corner of the house directly below my office, where the Causalis router and WiFi sits. My address is, shall we say, Rw House x (RwHx), which is also the street address of my housemate as well as of my APL. However, my housemate's APL RyHz is attached to a building whose street address is on a different road, namely Ry. The APLs are connected to a grey distributor box on the sidewalk of a third road some 50 metres away. A cable runs from that distributor to a point on the opposite sidewalk to my house on Rw, and runs beyond that further along Rw. There is a branch point opposite my house from which a cable runs past my house under my office window to the APL at RyHz (it also branches under my window to the neighbour's APL on RyHv).

Every so often, my housemate's Internet/phone connection has ceased to work. She has called the provider (BITel, a branch of the city utility company) to file a service difficulty. They check the line in real time, which is displayed, but they do not see her modem/router. This happens because, every so often, a Deutsche Telekom service engineer has come out to work on something at the distributor box, somehow has seen my housemate with registered street address RwHz connected somewhere else (namely RyHz Connector 7), and has then reconnected to APL-RwHx (usually Connector 2). I recall this happened three times in a particular eight-month period.

The service provider BITel contracts with Deutsche Telekom, as it must, for service using Deutsche Telekom's already-installed hardware. The “last mile” infrastructure up to and including the APLs belong to Deutsche Telekom, so Deutsche Telekom gets to do what it wants with them and theoretically neither BITel nor I are formally allowed to unscrew the APL lid to find out what is going on. But ... we have X-ray eyes. Besides, this is repeating behaviour.

The final time, we had three engineers out to check: Deutsche Telekom, then BITel, then Deutsche Telekom again. A Deutsche Telekom engineer came out to check the lines. He appears to have made one mistake, which is that he checked APL-RyHz and determined there was no connection to my housemate's router, whereas there is, on Connector 7. He went to my APL-RwHx, observed there was a live connection to Connector 2, connected his equipment, and determined that that was indeed the connection associated with my housemate's BITel account.

And now the point of this story. Deutsche Telekom contracts to provide a connection to an APL associated with a street address. The street address of my housemate is RwHx. That is where she is connected. For Deutsche Telekom, that connection and service is fulfilling its intended function. It follows (per IEV definition — better said, the “preferred” IEV definition) that it is reliable. As owner of the building, it is according to Deutsche Telekom my responsibility to provide a line from APL-RwHx to my housemate's apartment. I do provide a line, conformant with my responsibility, and APL-RwHx is *not* where that line is. That line goes to APL-RyHz, and has done since long before I owned the property. Since that line was not connected to the Deutsche Telekom network (that is, to the distribution box), according to BITel and my housemate and the contract, she is not receiving reliable service.

So who is right? Deutsche Telekom claims it is providing reliable service; BITel and my housemate claim it is not. The facts “on the ground” are the same for both, and both are right according to their reasoning. But what they claim is contradictory. Can this really be

so? Is reliability really so subjective? Do we really need paraconsistent[21] logic to describe the reality of daily life?

Obviously not. Deutsche Telekom and BITel/housemate have different interpretations of "intended function". Deutsche Telekom thinks "intended function" is providing a live line to the APL at the street address. BITel/housemate think "intended function" is providing a live line to the APL where my housemate's telephone line is connected.

This demonstrates that the concept of "intended function" must be unambiguous in order to have a workable concept of "reliability", namely one in which all engineers ascribe the same reliability properties to the exact same world state.

The story of course was resolved. Out came a BITel engineer to verify the topology; he then put in a request to Deutsche Telekom to connect my housemate's account back to APL-RyHz Connector 7; who sent an engineer the next day to do so. End of story. It's worked for a couple of years now.

3.5 Unambiguous Specification

There is an existing engineering term for "description of intended function such that all (rational, non-delusional, observation-capable) engineers ascribe the same meaning to that description in a given state of the world". It is *unambiguous specification*.

Furthermore, according to the observability condition elucidated above, that unambiguous specification shall involve only terms which are d/i-observable.

It is obvious in the story in Sub-section 3.4 that there are two descriptions of intended function, that these descriptions involve only d/i-observable phenomena (via the engineers' signal-generators and reception devices), and that these descriptions are different. There is no unambiguous d/i-observable specification of the telecommunications service. It follows that there is no way of ascertaining the reliability of that service.

From this observation, we can derive the principle:

> *7. Descriptions shall be, as far as possible, unambiguous d/i-observable specifications*

Where are we so far? We have devised seven principles, as well as some observations about the notion of reliability (as an example of what we take conceptual analysis for engineering to consist in).

The observations we have made about the notion of reliability are:

- There is engineered-kit, K, an artefact, to which we wish to ascribe a reliability, loosely an ability to function
- Such reliability is helpful as a value parameter: some kind of proportion of delivered-function to all actuations
- The value is dependent on a temporal parameter: number of demands (since new); hours of operation
- There must be an unambiguous d/i-observable specification of the function
- Various sorts are involved in an ascription of reliability to the kit K.

[21] Paraconsistent logic allows contradictions. In order not to be technically trivial, it must give up on the usual classical inference rule of "*ex falso quodlibet*": that from a contradiction, anything follows. In symbols, FROM (A && NOT A) INFER B, where A and B are arbitrary proposition symbols (called "propositional variables" in formal logic).

To determine an ability to function, one must be able to determine, for an observed behaviour within the range of behaviours of interest, whether it is consistent with the specification or not.

This leads to a further condition, on the kit K, that behaviour relevant to the function be d/i-observable, otherwise no comparison with a d/i-observable unambiguous specification of that function is possible. Also, there is a condition on the specification, that inference from the specification be relatively easily performed, for relevant behavioural characteristics. These conditions should be kept in mind when devising a definition of the term *reliability*. They follow as indicated from the loose conception *ability to function*, so they are strictly speaking part of the ConcAn of the term *reliability*. However, one could obviously formulate a general principle that:

> *8. As far as possible, where a term is defined in which an artefact is to be compared with a specification, the characteristics of behaviour coming within scope of the specification shall be d/i-observable, and the inferences from the specification concerning those characteristics shall be relatively easy.*

4 ConcAn Principles and the Definition of Reliability in IEV 192-01-24

4.1 The Principles

The principles elucidated in Section 3 are:

1. There is a vocabulary of architectural terms (for systems and parts of systems)
2. There is a vocabulary for descriptions of events, states, and behaviour
3. There is a basic vocabulary of sorts: at least:
    1. architectural concepts
    2. descriptions
    3. agents (human and artificial) and their organisation,
    4. environment-system relations and behaviour
    5. sociotechnical relations and behaviour
4. Use LSL, in particular to enable consistency-checking
5. Determine the arguments to (components of) a concept, their sorts, and thereby its arity
6. Use, as far as possible, d/i-observable terms
7. Descriptions shall be, as far as possible, unambiguous d/i-observable specifications
8. As far as possible, where a term is defined in which an artefact is to be compared with a specification, the characteristics of behaviour coming within scope of the specification shall be d/i-observable, and the inferences from the specification concerning those characteristics shall be relatively easy.

The definition of general *reliability* in the International Electrotechnical Vocabulary (IEV n. d.) is in Part 192, entry 192-01-24:

> *ability to perform as required, without failure, for a given time interval, under given conditions*

There are other definitions. The full list of definitions of reliability in the IEV is given in Appendix A. I shall consider variations below, after the initial analysis of 192-01-24.

4.2 Redefining Reliability Conformant with the Conditions

Let me call the terms *for a given time interval* and *under given conditions* the *auxiliary constraints*. Looking at the terms used, we have:

- ability
- perform
- as required
- failure
- the auxiliary constraints

According to Principle 5 of ConcAn, then, we need to establish what sorts are associated with these terms/phrases (including of course the auxiliary constraints).

We note that there is nothing about

- a measure
- a measured value
- (inferred parameter) the confidence in the measured value
- a specification
- what success or failure is
- what one can say about on-demand functions, for there seems to be no notion of *demand*, just that of *time interval*

There are some formal strictures required by the IEC terminology guidelines. *Reliability* is an *-ity*, a noun, its transcription as a definition must be more or less substitutive, that is, one should be able to replace any occurrence of the term *reliability* in an engineering assertion with its definition *ability to perform …* and retain the self-same meaning.

To analyse what is here being said, one might wish to broaden the syntactic categories used. Let us look at an adjectival form:

> Artefact K is *reliable* <in conditions, over a time interval> if and only if it is *able to perform* <in those conditions, over that time interval> <according to precept: namely, *as required, without failure*>

So ascribing an *ability* to K, as definition 192-01-24 seems to require, disappears when the syntactic form changes. The term *ability* is simply an artefact of the syntactic guidelines for IEC terminology.

Let us look instead at the phrase *able to perform*. When I say, “I am able to cycle into town now”, I do not mean, “I am cycling into town now”. I mean something like, “If called upon to do so, I will (successfully) cycle into town now”. Providing of course that my bike does not have a puncture, and that I don't have a stroke or a heart attack in the next little while (the auxiliary constraints). And I also don't mean “right now, this very second”, because I don't have my cycling shoes or helmet on, and my bike is still in the shed, locked up, so I would have to prepare, which takes a little time. So I mean <over a time interval>, the time interval being appropriate for the required preparation and execution of the action. So the inclusion of the auxiliary constraints <given conditions> and <given time interval> seem intuitively consonant with common usage of a term such as *able to perform*. We may conclude: the auxiliary constraints are appropriate for the intended use of the phrase *able to perform.*

What about the term *perform*? In common usage, it has something to do with expectation (of others). You *perform* when you are trying to execute some kind of behaviour of an

expected variety (by others). A play, a piece of music, a series of jokes at a microphone, taking part in a game. That expectation is implicit in 192-01-24, in the term *as required*: if something is required, surely there is a requirement?

It turns out we have all the terms to describe this already, in our basic vocabulary in the principles. *Able to perform as required, without failure* is nothing more or less than *behaves conformant with its specification*. (To get this to work, though, we need to presume here a condition on on-demand functions that a specification of an on-demand function allows dormant behaviour, namely that there is no specification-relevant state change when no demand is present[22].)

Using principles of ConcAn, namely 1, 2, and 3, we have succeeded in paraphrasing the IEV definition of *reliability*:

> *behaviour conformant with its specification, over a given time interval, under given conditions*

Let me call this definition *ConcAn-modified-192*. We have fulfilled Principle 5 in that we have basis sorts; an additional sort over the phrasing in 192-01-24, explicitly introduced, namely that of *specification*, which is also a basis sort, and whatever sorts are associated with the auxiliary constraints. We have the sorts:

- behaviour (a basis sort)
- specification (also a basis sort)
- the sorts of the auxiliary constraints

This definition is clearly Boolean. K is reliable (or not) under the auxiliary constraints.

The list of matters about which I noted above there was "nothing" explicit in the 192-01-24 definition can be revised. Specification and failure have been dealt with in the revised definition. The notion of time interval does not refer to what kind of behaviour (on-demand or continuous) which K exhibits, but rather the scope of the attribution of reliability, the period of time over which observations pertinent to reliability are made (this why this is part of the auxiliary conditions, but I still have to say what auxiliary conditions are).

4.3 Variations

We have:

- Reliability is an "ability" (most definitions) or a "probability" (192-05-05, 444-07-01, 448-12-05). These are fundamentally different notions, and thereby sorts. We have dealt with the term "ability" above; we shall consider "probability" below. What we can conclude in any case is that, even within IEV Section 192, "reliability" is a homonym, because two definitions of it attribute it to two different sorts. Homonyms are, however, excluded by IEC terminology guidelines. This is not an issue for ConcAn, though. Both ConcAn and SemAn are (I claim), *inter alia*, means of discovering synonyms and homonyms where they might be concealed. Here, that "reliability" exhibits homonyms is apparent on the surface.

[22] This is a principle of on-demand functionality, it is not a principle of terminology or ConcAn. We could aim to try to turn it into a terminological issue by using it explicitly in the definition of "*on-demand function*". But if we did that, we would need a further definition of "*intended-to-be on-demand function*" which we can call any function which reacts on demand, but which doesn't satisfy or cannot be shown to satisfy the statis-on-no-demand condition. Turning it into terminology does not avoid the issue that the condition is (meta-)physical, not terminological.

- How the "item" shall operate: "perform as required" or "perform a required function". One could discuss whether performing "as required" is broader in scope than performing a required function, for example, there might be additional properties of a performance which are desired/required other than that of executing a function. For example, that a programmable-electronic chip shall not heat up beyond a certain point when performing its required calculations. It is not strictly speaking a function that it shall not heat up, but we can certainly make such a requirement. I have analysed the definition using the possibly broader scope, and suggest this suffices for our purposes here.
- The auxiliary conditions are "given..." (most definitions), or they are "stated..." or "specified..." (the two definitions from Section 603). It would be very much a secondary issue to inquire whether these identifiers for means of expression all have the same meaning. More important is for the auxiliary conditions to be explicitly expressed. I do not pursue this further here.

The matters in the list which are not dealt with in the new ConcAn-modified-192, which (I claim) is semantically equivalent to that in 192-01-24, are *measure* and its *value*, thereby necessarily introducing a *confidence* in the value. This is the "probability" concept given in 192-05-05 and 444-07-01.

There is, of course, also a confidence involved in assessing whether behaviour conformed to specification under the auxiliary constraints — not everything may have been as practically observable as we might have wished, but this is not intrinsic to the definition, it is a consequence of empirical connection with the world in general. In contrast, the *confidence* involved in a measure of reliability is intrinsic to the use of the measure: for the self-same set of observations, there are different measure values associated with different confidence levels (in fact, an infinite set of pairs of <value, confidence>).

Conformant with the notion of "dispatch reliability", and indeed the notion of software reliability as construed by those who study it, as well as in other engineering contexts (Bedford and Cooke 2001, Section 3.2, p41), we might wish to say that reliability is rather something like:

> *over a given time interval, under fixed auxiliary conditions, the proportion of time (for continuous processes) or the proportion of demands (for on-demand processes) for which the ConcAn-modifed-192 is fulfilled, along with the confidence in this measure.*

Let me call this *measure-reliability*. Rather than being a Boolean value for given auxiliary conditions, as ConcAn-modified-192 is, the measure-reliability of K has type:

> (<auxiliary conditions>, <a proportion or percent>, <confidence>)

It should be clear that measure-reliability, which is just called "reliability" by statisticians and people who calculate it, is a pervasive concept in reliability engineering. Here a quote from a major textbook in the field (Birolini 2014, p1):

> *The expectation today is that complex equipment and systems are not only free from defects and systematic failures at time t = 0 (when they are put into operation), but also perform the required function failure free for a stated time interval and have a fail-safe behaviour in case of critical or catastrophic failures. However, the question of whether a given item will operate without failures during a stated period of time cannot be simply answered by yes or no, on the basis of a compliance test. Experience shows that only a probability for this occurrence can be given*

Indeed, measured-reliability is used in IEC standards. Tables 2 and 3 of IEC 61508-1:2010 give Safety Integrity Level requirements on random failures in terms (almost) of measure-reliability. What is missing in those tables is the notion of confidence. Over a given time interval, it is possible to be almost completely confident that <under auxiliary conditions> the function will execute correctly at least once, and at the same time have zero confidence that it will correctly execute all the time (respectively, on every demand). Indeed, this is the case with most software, and *ipso facto* with most devices which are software-based, which is the category of devices with which IEC 61508 is explicitly concerned. In the (misleading, and in part incorrect) explanation of statistical evaluation of the operational history of software in IEC 61508-7:2010, Annex D, Table D.1 gives measured-reliability conditions for both on-demand and continuous functionality, and attempts to relate it to safety integrity levels (SILs) — and also includes "confidence level". So it is clear that measured-reliability is relevant to electrotechnology, at least to safety. Indeed, a Software Safety Integrity Level is defined in IEC 61508-4:2010 Subclause 3.5.10 as a *measure of … confidence*…. So the notion of *confidence* specifically occurs. Here, again, is Birolini on the reliability of electronic components (Birolini 2014, Section 3.1.6 Reliability, p86):

> *The reliability of an electronic component can often be specified by its* failure rate λ. *Failure rate figures obtained from field data are valid if* intrinsic *failures can be* separated *from* extrinsic *ones and reliable data/information are available. Those figures given by component manufacturers are useful if calculated with appropriate values for the* activation energy…… *and* confidence level….

We may conclude that the definition of reliability in IEV 192-01-24 leaves a lot of questions open, and is not adequate to all the reliability concepts used in IEC standards, in particular those in IEC 61508. We might think that 192-05-05, measured-reliability, satisfies these additional needs: the probability of performing as required under the auxiliary conditions.

With "probability", though, we come up against Principle 6 of ConcAn. It should be d/i-observable. However, in any practical case we can only estimate it to a given level of confidence, as the quotations from Birolini (2014) indicate. The definition 192-05-05 gives no hint that, to fit any one given situation (collection of data), a variety of estimates of probability, each with a different confidence level, can be given. It may be disputed whether a definition is the correct place to say how a presentation of the item in question should look, but as a practical matter it could be hinted in a note that any given probability can only be an estimate, and needs to be accompanied by a confidence level.

This concludes this example of ConcAn (for now). It remains to say something about auxiliary constraints and the role they play in electrotechnical definitions, and thus in ConcAn.

5 Auxiliary Constraints — A Final Condition

Many concepts are general in that they apply to many sorts of thing. The concept of *reliability*, for example, applies to cars, software, and operator perceptions in a chemical plant. A statement of reliability of a car will be subject to certain constraints: the temperature outside, the road condition of the road or other surface on which it travels, the amount of water on the road (most cars do not do well in more than a few centimetres). A statement of reliability of a jet engine will include the flow of water it may ingest, as well as a mass of birds it might resiliently shred. A statement of reliability of software will

include none of those physical constraints. It may simply be conditioned on the ConcAn-modified-192 reliability of the hardware on which it is running. Reliability of operator perceptions will be conditioned on fatigue, biorhythmic parameters, perceptual capacity (acuity of eyesight, for example).

This leads to the notion of *auxiliary constraint*: conditions specific to the sort, as well as the branch of engineering, under which the term is to be used.

The reliability of a piece of kit K can only be made sense of when K is operational, so a general constraint is that one can only meaningfully talk of the reliability of K over at most its operational lifetime, and not beyond. This lifetime will likely include periods when K is not operational, for example during maintenance of K. During these periods, it also does not make sense to talk of the reliability of K[23].

Since the auxiliary constraints are dependent on the sort of the object to which the term is being applied, it does not necessarily make sense to list them when giving a general definition. However, it should be said that there do exist sort-specific auxiliary conditions for the definition to be applicable.

> *9. Where a definition is applicable to many different sorts, and the conditions under which the definition is applicable differ from sort to sort, the phrase "under sort-specific auxiliary constraints" or some equivalent ("under given conditions") should appear in the definition*

6 Matters Not Dealt With in (Current) ConcAn

There are some important issues in terminology which are not addressed, or not addressed adequately, by the ConcAn principles that have been formulated here. Here are some.

6.1 Goal Use Versus Actual Use

When defining a concept, the definer usually has some idea of why, wherefore, and how she intends the term to be used. This may or may not reflect the way in which electrotechnologists actually use the term (if it is already in use). Indeed, if the term is in use, it may well have different meanings to different users, in which case fixing on even one of those meanings, which is what a terminological definition does, inevitably alienates all other uses[24].

If those uses "in the field" are unknown, then they can sometimes be detected using machine-learning techniques on restricted corpora (namely, engineering documents in use) (Arndt and Schnäpp 2018). This is a key result of Project Harbsafe[25] (within which the initial work on ConcAn was performed). Two issues in applying such techniques are:

- the definition of a large-enough domain-specific corpus (Project Harbsafe used selected IEC documents in cybersecurity and safety; it is not clear at time of writing which corpus size is "large enough")

[23] It may be, though, that the reliability of the overall system S of which K is part is important during periods in which K is not operational (is undergoing maintenance, for example). So it may be important for the reliability of S that the function of K is substituted for during K's "down time"

[24] …and users.

[25] See Acknowledgments, below

- the acquisition of enough volunteers to render judgements which enable reliable results from the machine-learning techniques

ConcAn does not employ such techniques; it is intellectual/symbolic analysis.

Sometimes, divergent use "in the field" is known. The authors of IEC 61511:2016, the process-industry-specific specialisation of IEC 61508, have noted the following divergences of their terminology from that of IEC 61508, explicitly because of common usage in their field. Many of these are key terms!

- 3.2.1 architecture
- 3.2.26 functional safety assessment.
- 3.2.29 hardware safety integrity.
- 3.2.56 process risk. A synonym for "EUC risk" in IEC 61508.
- 3.2.57 programmable electronics.
- 3.2.64 redundancy
- 3.2.67 safe state
- 3.2.69 safety function
- 3.2.73 safety integrity level
- 3.2.75 safety manual
- 3.2.79 software
- 3.2.84 system
- 3.2.91.1 type A device
- 3.2.91.2 type B device

Technically, these terms are thereby homonyms in the IEC corpus. As noted above, IEC terminology guidelines say that homonyms shall be avoided: these homonymic variant definitions are intentional, and occur within the responsibilities of one IEC Subcommittee, SC65A. ConcAn currently implements no proposals or guidelines for how to deal with deviant usage based on practicality.

6.2 Coherence

ConcAn recommends the use of LSL, and (semi-)automated proof checking, to check for consistency of a definition/mutual consistency of groups of definitions. But currently ConcAn includes no guidelines as to how to pick candidate (sets of) definitions for consistency analysis. It is left to the analyst's thoroughness and judgement.

6.3 Efficacy and Coverage

A key question in devising terminology is whether all technical questions related to a concept (or set of concepts) allow themselves to be concisely expressed using the concept (set of concepts). For example, I had observed confusions generated in discussion through and about use of the technical definitions concerning harm, risk and safety in IEC 61508 and devised my own set of definitions which I felt avoided the problems (Ladkin 2008).

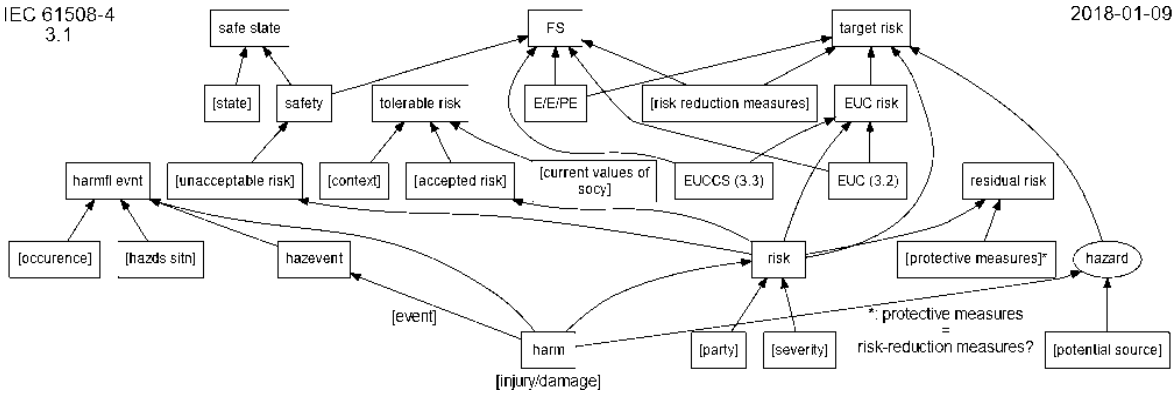

Figure 2 ~ Semantic Dependency Graph of the Terminology in IEC 61508-4 Subclause 3.1

Semantic dependency graphs (SDGs) indicate a more efficient use of terminology. These are discrete directed graphs (composed of nodes and directed edges between them) in which a node represents a term, and a directed edge from the node labelled A to the node labelled B indicates that the term "A" occurs in the definiens of term "B". Figure 2 shows the semantic dependencies in this terminology in IEC 61508-4 Subclause 3.1.

Figure 3 shows the semantic dependencies amongst the similar terms in Ladkin (2008), with some terms grouped (a "high-level" SDG).

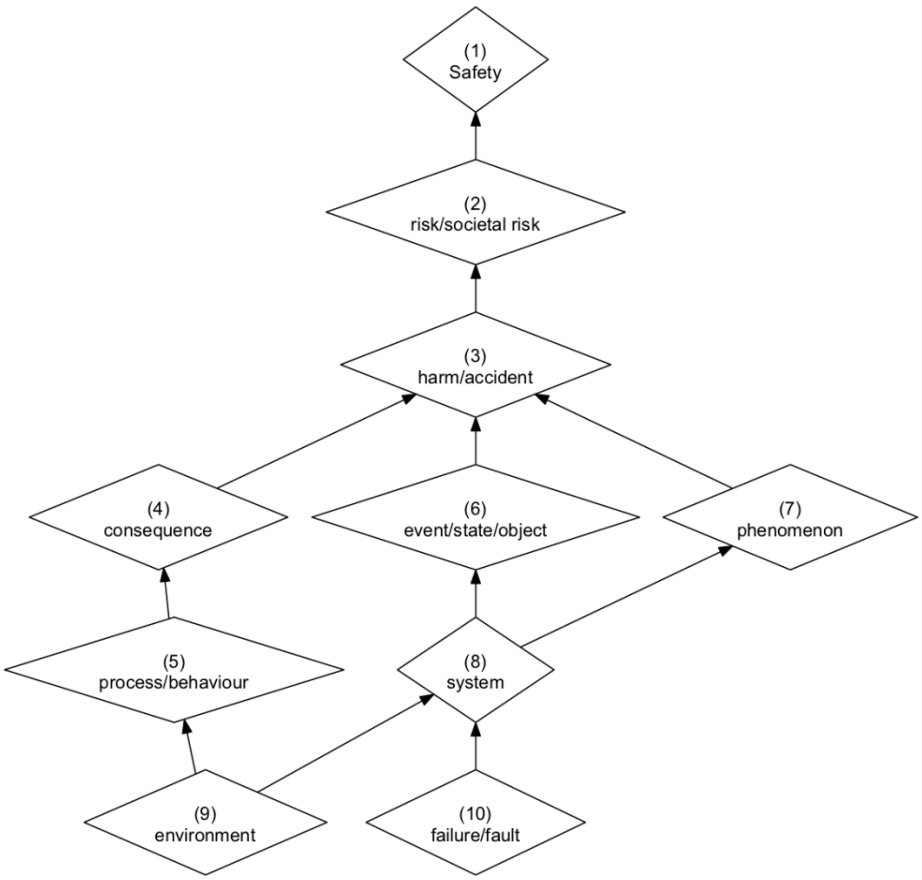

Figure 3 ~ High-level SDG of the Vocabulary of Ladkin (2008)

ConcAn currently provides no mechanisms or guidelines for assessing efficacy or coverage.

7 Summary

Conceptual analysis is a term from philosophy with varied meanings. I have argued that some of the varied techniques are applicable to electrotechnological vocabulary, and I call this adaptation ConcAn. I have illustrated ConcAn most simply on a specific IEV term, 151-11-03 *electric*, and devised a collection of principles under which ConcAn proceeds by considering the IEV term 192-01-24 *reliability*. These principles (so far) are:

1. There is a vocabulary of architectural terms (for systems and parts of systems)
2. There is a vocabulary for descriptions of events, states, and behaviour
3. There is a basic vocabulary of sorts: at least

    1. architectural concepts
    2. descriptions
    3. agents (human and artificial) and their organisation,
    4. environment-system relations and behaviour
    5. sociotechnical relations and behaviour

4. Use LSL, in particular to enable consistency-checking
5. Determine the arguments to (components of) a concept, their sorts, and thereby its arity
6. Use, as far as possible, d/i-observable terms
7. Descriptions shall be, as far as possible, unambiguous d/i-observable specifications
8. As far as possible, where a term is defined in which an artefact is to be compared with a specification, the characteristics of behaviour coming within scope of the specification shall be d/i-observable, and the inferences from the specification concerning those characteristics shall be relatively easy.
9. Where a definition is applicable to many different sorts, and the conditions under which the definition is applicable differ from sort to sort, the phrase “under sort-specific auxiliary constraints” or some equivalent (“under given conditions”) should appear in the definition

~

Correspondence Address

Corresponding e-mail address: ladkin@causalis.com.

Acknowledgments

The initial work on ConcAn was performed in the project Harbsafe, financed by (as it then was) the German Federal Ministry for Economic Affairs and Energy, No. 03TNG006A-B in the Wipano programme, awarded to the Technical University of Braunschweig (TU-BS), Institut IVA, and DKE (the Deutsche Kommission Elektrotechnik Elektronik Informationstechnik im DIN und VDE), which is a German electrotechnical standardisation organisation, in 2017—2019. Causalis Ingenieurgesellschaft worked in Harbsafe as a subcontractor to the Technical University of Braunschweig.

References

Arndt S. and Schnäpp D. (2018). *Harbsafe-162 — A Domain-Specific Data Set for the Intrinsic Evaluation of Semantic Representations for Terminological Data*. Preprint submitted for publication, Technical University of Braunschweig IVA — Institute for Traffic Safety and Automation Engineering, 2018.

Beaney M. (2003). *Analysis*. In: Stanford Encyclopedia of Philosophy. 2003, revised 2014. Available from https://plato.stanford.edu/entries/analysis/. Accessed 27th July 2023.

Bedford T. and Cooke R. (2001). *Probabilistic Risk Analysis*. Cambridge University Press, Cambridge.

Birolini A. (2014). *Reliability Engineering: Theory and Practice*, Seventh Edition, Springer-Verlag, London.

Blackburn S. (1994). *The Oxford Dictionary of Philosophy*. Oxford University Press, Oxford.

Carlson C. S. (2012). *Effective FMEAs: Achieving Safe, Reliable, and Economical Products and Processes using Failure Mode and Effects Analysis*. John Wiley & Sons, Hoboken NJ.

Collins J., Hall N., and Paul L. A. (editors). (2004). *Causation and Counterfactuals*. MIT Press, Cambridge, MA.

Fodor J. (2004). *Water's water everywhere; Review of Hughes on Kripke*. London Review of Books 26(20), 21 October 2004. Available from https://www.lrb.co.uk/v26/n20/jerry-fodor/waters-water-everywhere. Accessed 27th July 2023.

Galton A. (1996). *Note on a Lemma of Ladkin*. J. Logic and Computation 6(1):1-4, February 1996.

IEC61508. (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems*. IEC 61508, in 7 parts, 2nd Edition, 2010. International Electrotechnical Commission, Geneva.

IEC61511. (2016). *Functional safety — Safety instrumented systems for the process industry sector — Part 1: Framework, definitions, system, hardware and application programming requirements*. IEC 61511-1, 2nd Edition, 2016. International Electrotechnical Commission, Geneva.

IEC TR 63069. (2019). *Industrial-process measurement, control and automation - Framework for functional safety and security*. IEC TR 63069:2019. International Electrotechnical Commission, Geneva.

IEC Glossary. (no date). *IEC Glossary*. International Electrotechnical Commission. Available from https://std.iec.ch/glossary. Accessed 29th July 2023. In particular, the entry for "reliability" as in 191-02-06 is available at https://std.iec.ch/terms/terms.nsf/3385f156e728849bc1256e8c00278ad2/b73f2f0202273857c1257c46004f0134?OpenDocument. (Note that there are many other entries for "reliability", and that this particular entry may not be found by using the internal searching techniques of the Glossary.)

IEV. (no date). *International Electrotechnical Vocabulary*. International Electrotechnical Commission, IEC 60050. Available from https://www.electropedia.org. Accessed 29th July 2023.

Jackson M. A. (2013). *Topsy-Turvy Requirements*. In: Meyer B. (editor). *Modelling and Quality in Requirements Engineering : Essays dedicated to Martin Glinz on the occasion of his 60th birthday*. Verlag-Haus Monsenstein u. Vannerdat, Münster.

Kripke S. (1980). *Naming and Necessity*. Blackwell, Oxford.

Ladkin P. (1987a). *Models of Axioms for Time Intervals*. Proceedings of AAAI-87. The Association for the Advancement of Artificial Intelligence. Available at https://www.aaai.org/Papers/AAAI/1987/AAAI87-042.pdf. Accessed 29th July 2023.

Ladkin P. B. (1987b). *The Logic of Time Representation*. Ph.D. Thesis, Group in Logic and the Methodology of Science, University of California, Berkeley. Available from https://www.researchgate.net/publication/2242666_The_Logic_of_Time_Representation. Accessed 29th July 2023.

Ladkin P. B. (2001). *Causal System Analysis*. RVS Group, Bielefeld University. Available from https://rvs-bi.de/publications/books/CausalSystemAnalysis/index.html. Draft e-textbook accessed 29th July 2023.

Ladkin P. B. (2008). *Definitions for Safety Engineering*. Causalis Limited. Available from https://causalis.com/90-publications/99-downloads/DefinitionsForSafetyEngineering.pdf. Accessed 29th July 2023.

Ladkin P. B. (2017). *Digital System Safety*. RVS Group, Bielefeld University. Available from https://rvs-bi.de/publications/RVS-Bk-17-02.html. Draft e-textbook accessed 29th July 2023.

Ladkin P. B. (2019). *Architectural Concepts in Key TC 65 Standards on Safety and Cybersecurity*. Project Harbsafe Working Paper, Version 1 of 2019-05-14, Causalis Ingenieurgesellschaft mbH.

Ladkin P. B. (2020). *IEC TR 63069, Security Environments and Security-Risk Analysis*. In: Parsons M., Nicholson M. (editors) *Assuring Safe Autonomy, Proceedings of the Twenty-eighth Safety-Critical Systems Symposium, York, UK*. SCSC-154, SCSC C.I.C. 2020.

Ladkin P. B., Lou X., and Schnäpp D. (2023). *The Terminological Analysis Method SemAn and its Implementation*. Safety-Critical Systems eJournal 2(1). SCSC-183. Safety-Critical Systems Club. Available from: https://scsc.uk/r183.4:1#page=55. Accessed 29th July 2023.

Lewis D. K. (1973a). *Causation*. Journal of Philosophy 70(17):556-567, 1973. Reprinted with postscripts in (Lewis 1986).

Lewis D. K. (1973b). *Counterfactuals*. Blackwell, Oxford, 1973, reissued 2001.

Lewis, D. K. (1986). *Philosophical Papers* (Vol. II). Oxford University Press, Oxford.

Littlewood B. & Strigini L. (1993). *Validation of Ultrahigh Dependability for Software-Based Systems.* Communications of the ACM (CACM), 36(11), pp. 69–80. Available from https://openaccess.city.ac.uk/id/eprint/1251/1/CACMnov93.pdf. Accessed 29th July 2023.

Magidor O. (2019). *Category Mistakes*. In: Zalta E. N., Nodelman U. (editors) *Stanford Encyclopedia of Philosophy, Fall 2022 Edition*. https://plato.stanford.edu/archives/fall2022/entries/category-mistakes. Accessed 14th November 2023.

Margolis E. and Laurence S. (2005). Concepts. In: Stanford Encyclopedia of Philosophy. 2005, revised 2019. (Section 5 talks specifically about conceptual analysis). Available from https://plato.stanford.edu/entries/concepts/. Accessed 27th July 2023.

Neyman J. (1937). *Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability*. Philosophical Transactions of the Royal Society of London A, 236:333–380. 30th August 1937.

Paul L. A. and Hall N. (2013). *Causation: A User's Guide*. Oxford University Press, Oxford.

Quine W. V. O. (1951). *Two Dogmas of Empiricism*. The Philosophical Review 60(1):20-43, Jan., 1951.

Quine W. V. O. (1960). *Word and Object*. The M.I.T Press, Cambridge, MA.

Ricque B. (2019). Bertrand Ricque: personal communications with the author, various dates in 2019.

Riemenschneider R. (1984). Robert Riemenschneider: personal communications with the author, in Berkeley, CA and environs, 1976-1984

Robinson arithmetic. (2023). In Wikipedia. Accessed 29th July 2023. https://en.wikipedia.org/wiki/Robinson_arithmetic

Spafford E. H., Metcalf L., and Dykstra J. (2023). *Cybersecurity Myths and Misconceptions*. Addison-Wesley/Pearson Education, Boston MA.

Appendix A. Definitions of "Reliability" in the On-Line International Electrotechnical Vocabulary

The definitions here are all those in the on-line version of the International Electrotechnical Vocabulary (IEV n.d.) which define "reliability" of something or other. It does not include terms which use the word "reliability" only as part of their designation, for instance "reliability block diagrams" (which are diagrams) or "reliability model" (which is a probabilistic model).

There are some minor syntactic violations of IEC guidelines in these definitions; in 395-07-131, 448-12-05, 603-05-01, 603-05-02, the initial "the" should not appear.

In 312-07-06, the definition is said to be taken from 191-02-06. This refers to the previous IEC 60050 Section 191: Dependability and quality of service. According to the IEC Webstore, this section was replaced in 2015 by Section 192: Dependability.

The numerical designation of a definition is threefold. The first number refers to the electrotechnical branch classification, and associates the definition with a specific Technical Committee of the IEC. The second number relates to that technical committee's classification of its areas of concern. The third number is just a sequence number of a series of definitions. The main branch and the subclassification are given after the definition number below.

Each definition has three parts:

- Number, along with branch and subclassification. Number is shown in bold-face font.
- Definiendum: the term which is defined (this may include qualifications, in angle brackets "<...>").
- Definiens: the definition, which is mainly meant to be substitutive. "Substitutive" means that the definiens may substitute for the definiendum in any syntactic context while retaining the exact same meaning. This may or may not be literally the case here, but substitutional definition is a well-researched area of philosophical logic and I suggest it is more or less clear what the intention is, even if it may be imperfectly realised in the IEV.

192-01-24 Dependability / Basic concepts

reliability <of an item>

> ability to perform as required, without failure, for a given time interval, under given conditions

Note 1 to entry: The time interval duration can be expressed in units appropriate to the item concerned, e.g. calendar time, operating cycles, distance run, etc., and the units should always be clearly stated.

Note 2 to entry: Given conditions include aspects that affect reliability, such as: mode of operation, stress levels, environmental conditions, and maintenance.

Note 3 to entry: Reliability can be quantified using measures defined in Section 192-05, Reliability related concepts: measures

192-05-05 Dependability / Reliability related concepts: measures

reliability <measure>

> probability of performing as required for the time interval (t1, t2), under given conditions

Note 1 to entry: Given conditions include aspects that affect reliability, such as: mode of operation; stress levels; environmental conditions; and maintenance, where applicable.

Note 2 to entry: It is usually assumed that the item is in a state to perform as required at the beginning of the time interval.

Note 3 to entry: When t1 = 0 and t2 = t, then R(0, t) is denoted simply as R(t) and termed the reliability function, or survival function of the item. See IEC 61703, Mathematical expressions for reliability, availability, maintainability and maintenance support terms, for more details.

Note 4 to entry: See also reliability, <of an item> (192-01-24).

312-07-06 Electrical and electronic measurements — General terms relating to electrical measurements / Performance

reliability (performance)

> ability of an item to perform a required function under given conditions for a given time interval

[SOURCE: 191-02-06]

(PBL Note: see preliminary comments to this Appendix. Section 191 no longer exists in the IEV. The definition 191-02-06 does exist, though, in the IEC Glossary, which contains the following entry for "reliability", inter alia:

> *ability of an item to perform a required function under given conditions for a given time interval*
> *NOTE 1 It is generally assumed that the item is in a state to perform this required function at the beginning of the time interval.*
> *NOTE 2 The term "reliability" is also used as a measure of reliability performance (see IEV 191-12-01).*
> *[IEV191-02-06]*

The notes here are important, in particular NOTE 1, which however is not adopted in 312-07-06. The References entry for the IEC Glossary (n.d.) contains the precise URL for this entry in the glossary.)

395-07-131 Nuclear instrumentation: Physical phenomena, basic concepts, instruments, systems, equipment and detectors / Nuclear fission reactors, including the nuclear fuel cycle and thermonuclear facilities

reliability

> the ability of an item to perform a required function under given conditions for a given time interval

Note 1 to entry: It is generally assumed that the item is in a state to perform this required function at the beginning of the time interval.

Note 2 to entry: Generally, reliability performance is quantified using appropriate measures. In some applications, these measures include an expression of reliability performance as a probability, which is also called reliability.

444-07-01 Elementary relays / Endurance

relay reliability

> probability that a relay can perform a required function under given conditions for a given duration or number of cycles

Note — It is assumed that the relay is able to perform this required function in its initial condition.

448-12-05 Power system protection / Reliability of protection

reliability of protection

> the probability that a protection can perform a required function under given conditions for a given time interval

Note — The required function for protection is to operate when required to do so and not to operate when not required to do so.

603-05-01 Generation, transmission and distribution of electricity — Power systems planning and managements / Power system reliability

reliability of an item

> the ability of an item to perform a required function under stated conditions for a specified period of time

603-05-02 Generation, transmission and distribution of electricity — Power systems planning and managements / Power system reliability

service reliability

> the ability of a power system to meet its supply function under stated conditions for a specified period of time

692-01-13 Generation, transmission and distribution of electrical energy — Dependability and quality of service of electric power systems / System concepts

reliability <of an item>

> See 191-01-24.

692-01-14 Generation, transmission and distribution of electrical energy — Dependability and quality of service of electric power systems / System concepts

service reliability

> ability to adequately satisfy the demand under given operating conditions for a given time interval…

This collation page left blank intentionally.

ISSN 2754-1118 (Online)
ISSN 2753-6599 (Print)

The Open Autonomy Safety Case Framework

Michael Wagner and Carmen Carlan

Edge Case Research Inc., Pittsburgh, United States

Abstract

A system safety case is a compelling, comprehensible, and valid argument about the satisfaction of the safety goals of a given system operating in a given environment supported by convincing evidence. Since the publication of UL 4600 in 2020, safety cases have become a best practice for measuring, managing, and communicating the safety of autonomous vehicles (AVs). Although UL 4600 provides guidance on how to build the safety case for an AV, the complexity of AVs and their operating environments, the novelty of the used technology, the need for complying with various regulations and technical standards, and for addressing cybersecurity concerns and ethical considerations make the development of safety cases for AVs challenging. To this end, safety case frameworks have been proposed that bring strategies, argument templates, and other guidance together to support the development of a safety case. This paper introduces the Open Autonomy Safety Case Framework, developed over years of work with the autonomous vehicle industry, as a roadmap for how AVs can be deployed safely and responsibly.

1 Introduction

The desire to use safety cases as safety communication concepts for commercial autonomous vehicles (AVs) is increasing, especially after the publication of UL 4600. A safety case is defined by UL 4600 as a "*structured argument, supported by a body of evidence, that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given environment*". Arguments in safety cases are usually inductive, meaning that the truth of its premises provides some grounds for its conclusion. Safety case approaches support the flexibility needed by safety processes for autonomous vehicles to address novel technologies and utilise emerging safety strategies without prescriptive guidance from standards or accepted practices (Koopman et al. 2019).

AVs are systems consisting of components implementing complex algorithms, sensors, and machine learning (ML) components. The development and management of safety cases for AVs is challenging. A safety case for an AV shall argue about the confidence that all relevant mishaps, the causing hazards, and the causal factors of the hazards are identified and how the implementation of mitigations is sufficient to mitigate the risk of the identified hazard. To this end, a correct understanding of the behaviour and performance of such components and their interaction is needed. Whereas a safety case is an argument about the safety of a system in a given operating environment, AVs will operate in real-world environments, which are complex and unpredictable. This means that the validity and even soundness of the safety case may change over time. Arguments in safety cases are valid if the premises provide enough confidence that the conclusions are true. Valid arguments are sound when the premises are true. The development of AVs often uses novel technology, which is yet to be standardised. Developing compelling

arguments supported by evidence generated using novel technology may impose specific challenges. Further, the safety case of AVs shall demonstrate compliance with a large set of standards and regulations. Lastly, a safety case of an AV shall also discuss safety-related cybersecurity and ethical considerations.

Safety case frameworks facilitate the creation of safety cases and guide their management. A safety case framework (SCF) captures the broad guidance necessary to develop a safe system and an adequate safety case. Apart from a templated argument structure, an SCF entails supporting concepts and strategies that guide the expansion of the argument, templates for supporting evidence, process definitions, and requirements checklists.

This paper introduces the Open Autonomy Safety Case Framework (OASCF), developed within Edge Case Research (ECR) over years of work with the autonomous vehicle industry, as a roadmap for how you can deploy safely and responsibly. Further, OASCF draws on Edge Case's experience writing and applying standards such as UL 4600, MIL-STD-882E, ISO 21448:2022, and ISO 26262:2018, as well as supporting developers of safe autonomous trucks and cars and assessing the safety of complex defence systems.

The OASCF is an overarching strategy and development philosophy for complex autonomous systems in the automotive industry. The goal of OASCF is to support safety engineers in developing a safety case in compliance with UL 4600 that argues that the system is developed in compliance with relevant standards mentioned by UL 4600, such as ISO 26262 and ISO 21448. Whereas the framework scopes automotive systems, UL 4600 also recommends following best practices from other industries, such as aerospace (Federal Aviation Administration (FAA) documents) and military (MIL-STD-882E). OASCF establishes how to implement a safety management system, communicate risks to stakeholders, and navigate relevant standards and regulations. The framework also enables effective, independent assessment processes defined in UL 4600. OASCF is *open*, meaning that it is implementation agnostic. OASCF can be used in different projects, implementing AVs for different use cases. Further, its openness is given by the fact that, hereby, we make the first nine levels of the templated high-level argumentation strategy public, because we believe that safety should be a prerequisite for deploying AVs, and not a competitive advantage. Therefore, we want to support the AV community to take the first steps in planning their safety programs.

The goals of OASCF are to:

- Enable deployment of safer autonomy by providing adequate safety management processes that developers can quickly adopt;
- Engender trust in safe autonomy by enabling AV developers to have a fully transparent safety case backed by clear evidence and a conformance monitoring plan;
- Drive effective regulation of autonomous vehicles; and
- Accelerate insurance and risk rating for fleets of autonomous vehicles.

2 Background on Safety Cases

The concept of safety cases was introduced thirty decades ago when, in 1989, the Control of the Industrial Major Accident Hazards in the British chemical industry mandated the generation of a written report named *safety case*, arguing about the mitigation of the hazards and risks of a site.

A *system safety case* is a well-constructed, easily understandable, and valid argument structured around ensuring the fulfilment of safety goals for a specific system operating within a defined environment. The safety goals are measures to mitigate the risk associated to the identified system hazards. The argument within the safety case is substantiated by compelling evidence. From a structural standpoint, a safety case comprises three fundamental components: safety claims regarding the system in question, a compilation of supporting evidence (such as safety analyses, software inspections, or functional tests) generated throughout the safety engineering lifecycle, and a rationale outlining how the available evidence contributes to demonstrating the achievement of the safety assertions. The claims can be further subdivided into sub-claims until each sub-claim is directly substantiated by evidence. The argument systematically captures the logical connections between claims, sub-claims, and evidence. The evidence supporting the argument is typically diverse, encompassing both quantitative and qualitative aspects and analytical and empirical data. Structured arguments may be graphically represented using different languages. One such language is the Goal Structuring Notation (GSN). Figure 1 shows an exemplar structured argument graphically represented using the Goal Structuring Notation (GSN). In GSN, the safety claims are represented via goals, the evidence via solution elements, and the rationale for how the safety claims are supported by evidence via strategies.

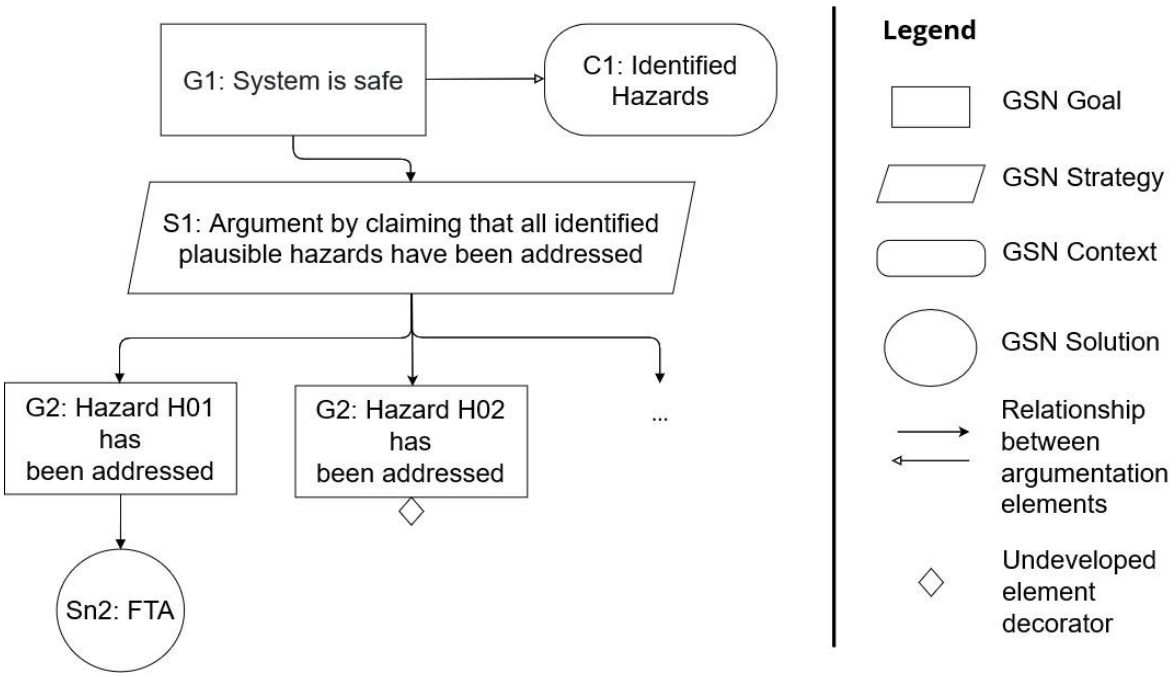

Figure 1 ~ Exemplar Structured Argument in GSN

Irrespective of the application domain, a safety case encompasses details about the system under evaluation and its operational environment. This includes defining the acceptable risks, identifying hazards, assessing the associated risks, specifying the adopted safety management system, and outlining responsibilities and the organisational safety policy.

In the automotive industry, Part 2 of ISO 26262 mandates that the system safety case compiles the system artefacts generated throughout the safety life cycle. Additionally, Part 10 of ISO 26262 mandates the presence of a plan for safety case maintenance in response to system changes. ISO 21448 offers examples of structured arguments concerning the achievement of Safety of the Intended Functionality (SOTIF). Whereas the Motor Industry Research Association's (MISRA's) Development Guidelines for Automotive

Safety Arguments (MISRA 2019), a standard for software development in automotive systems, offers guidelines for articulating arguments related to software correctness.

UL 4600 outlines a set of best practices and considerations for developing system safety cases for autonomous systems. It supports safety engineers by prompting critical questions such as, 'Have you considered this?'. Moreover, UL 4600 emphasises the importance of analysing the impact of system changes on safety cases.

3 Live Safety Cases

It is not enough to build a single, static safety case that quickly becomes out of date once approved. Safety arguments that do not accurately reflect the current state of the system give a false sense of system safety. Safety cases should be considered living documents updated, reviewed, and published after each safety-relevant change to the system.

In compliance with UL 4600, the OASCF is based on continuous feedback loops to resolve sources of aleatoric and epistemic uncertainty in residual risk claims. Many of these feedback loops involve conventional incident reporting followed by hazard analysis and mitigation using level-of-rigour activities. Resolving uncertainty is especially important for applications of autonomous vehicles in open-ended use cases in the real world. This fact is compounded by novel technologies such as machine learning. The safety cases created with OASCF are intended to be live (see Figure 2).

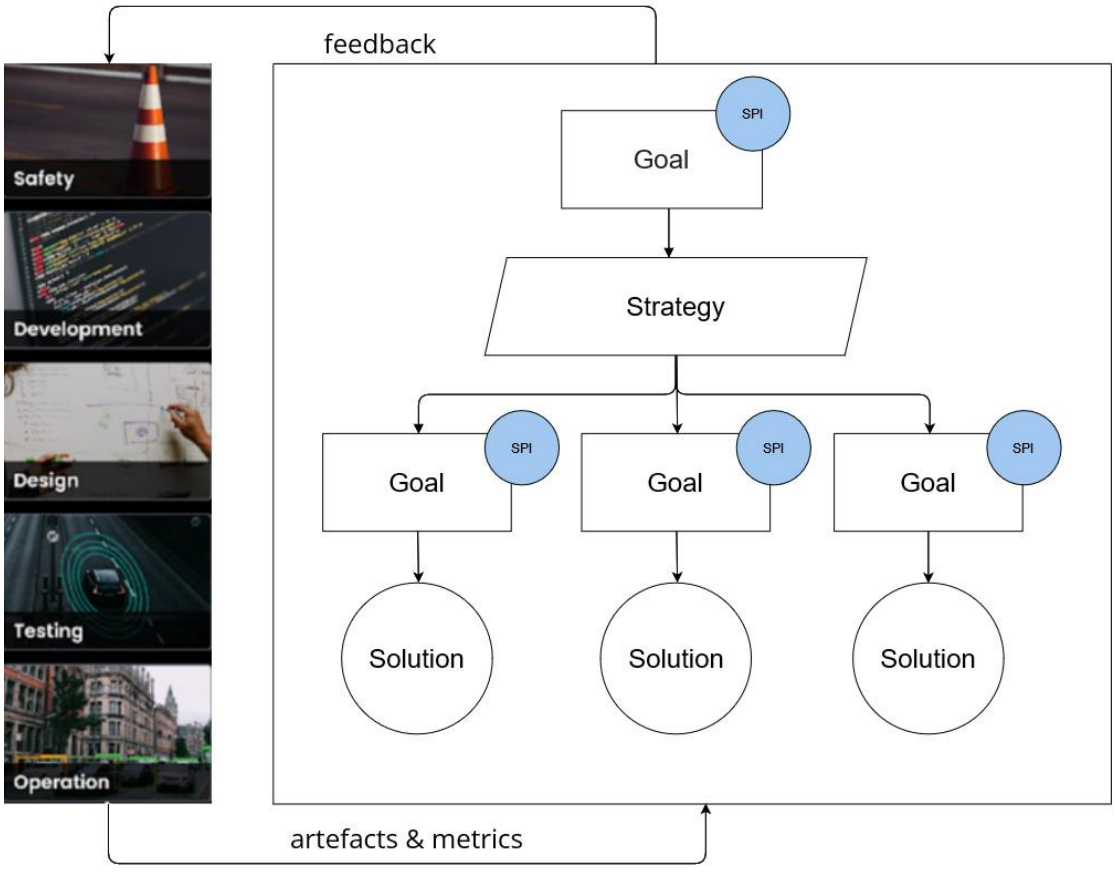

Figure 2 ~ Overview of the concept of live safety cases, which have dynamic links to system artefacts, and which have SPIs attached to the claims to evaluate their soundness

In comparison with regular safety cases, live safety cases have a set of additional properties:

- The truth of the claims in live safety cases may be automatically checked based on the existing evidence;
- Live safety cases are continuously fed with the most current safety evidence; and
- Live safety cases are sensitive to system changes. Whenever a system change occurs, the impact of the change on the safety case is automatically detected.

To this end, the claims in our OASCF are tied to templated Safety Performance Indicators (SPIs). While evidence and SPIs within the OASCF involve metrics data and thresholds, their purposes differ. Evidence supports the truth of a claim and is generally collected prior to an assessment. In contrast, SPIs detect situations where claims might be false. They are central to an assurance case monitoring plan defined by UL 4600. Templated SPIs can be tailored to data sources tied to a user's functional architecture and safety management system implementation.

Behavioural SPIs trace to autonomy functions such as perception, prediction, planning, and control. Operational SPIs trace to safety management systems, processes, and safety culture. SPIs are of two types: lagging and leading. Lagging SPIs directly measure the residual risk during system operation. Examples of lagging SPIs include fatalities, other crashes, violations of traffic rules, and near hits. Lagging SPIs are associated with claims higher up in the safety case and are measured during operation. Confidence in lagging SPIs increases as the deployment scales and, consequently, the volume of available data grows. Leading SPIs are measured both earlier in the system development lifecycle, including during simulation and road testing, and during deployment and operation, and may be used to predict system safety. Examples of leading SPIs include malfunctions, failures of system components, rates of sensor malfunctioning, perception failure rates, and software component execution faults. Leading SPIs are associated with claims deeper in the safety case. Leading SPIs may be used to identify so-called "triggering conditions" specified in ISO 21448, which identify operational situations in which an autonomy algorithm fails to mitigate hazards as intended. Figure 3 shows a fault tree that models how the occurrence of triggering conditions can lead to a loss event, such as a collision with a pedestrian. Confidence in the predictive nature of SPIs is to be gained by continuously comparing the predicted data with the actual data.

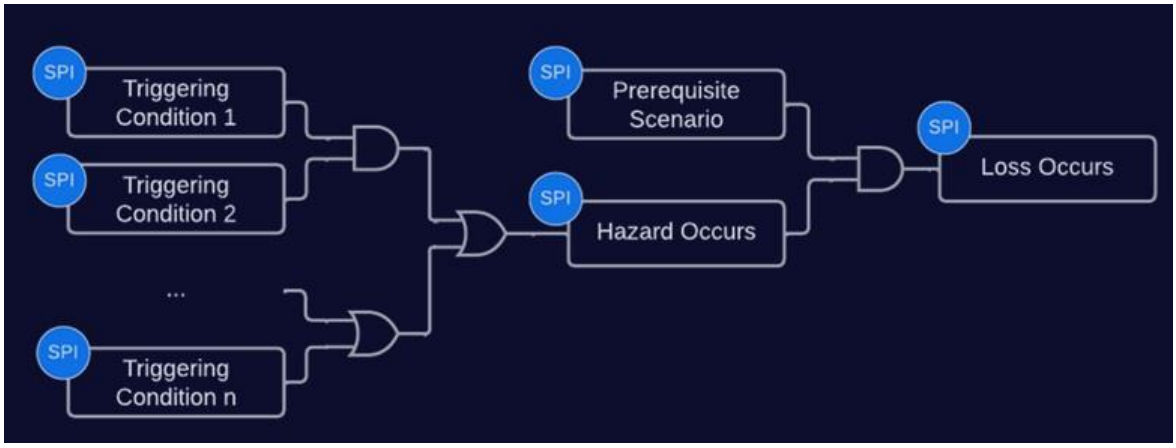

Figure 3 ~ SOTIF Fault Tree Annotated with SPIs

During system development, SPIs may detect the hazards whose desired risk reduction has not yet been achieved via the system design. One countermeasure could be to apply

procedural mitigations to allow system maturation while maintaining safety. At the transition into operational deployment, all SPIs shall be satisfied. However, because of a lack of sufficient data, the confidence in the satisfaction of the SPIs is low. Post-deployment monitoring of SPIs aims to increase this confidence. Also, post-deployment, SPIs can detect violations of claims and assumptions in the safety case that invalidate the argument. Violations of the safety case detected in operation should trigger remediation processes, including potentially grounding the fleet while the causes are diagnosed, and the system or safety argument is updated.

Arguments in live safety cases shall be supported by evidence system artefacts, such as simulation test results, feedback from fleet vehicles, and dynamic hazard tracking mitigations referenced via dynamic links. Dynamic links allow the safety case to always reference the latest version of the system artefacts.

Given changes in the referenced evidence, the impact of those changes on the safety arguments needs to be analysed. Live safety cases shall also support automated safety argument change impact analysis (CIA) and shall be structured in a manner that reduces the impact of foreseeable changes. Given a change in a part of a safety case or in a referenced system artefact, the impact of the change on the elements of the safety case shall be automatically identified. Several approaches for automated safety argument CIA have been proposed in the literature. In Cârlan et al. (2022), we proposed an approach for a sound, automated safety argument CIA based on semantically rich annotations of dynamic links between argument elements and system artefacts, and of traces between argumentation elements.

Live safety cases have many benefits, including:

- Tracking and communicating development progress internally;
- Confidently communicating residual risk externally;
- Automating data flows with standard and custom data connectors;
- Creating and managing interdependencies between complex sets of claims;
- Tracking hazards, mitigations, and requirements sets;
- Monitoring for SPI violations;
- Providing continuous risk management updates; and
- Identifying edge cases that need to be addressed.

4 The Proposed Safety Argumentation Strategy

4.1 Preface

Our proposed OASCF provides its user with an overarching strategy and development philosophy toward a safe system and a compelling safety case including the following:

- A safety assurance process based on relevant safety standards;
- A set of techniques and templates for the generation of system development artefacts that are to be used as safety evidence and checklists for their assessment;
- A templated top-level argumentation strategy for why the system is safe, based on the proposed safety process, from which ANSI/UL 4600 compliant safety cases can be built;

- Assessment criteria for a well-formed and sound safety case based on the state of the art and best practices, such as ISO/IEC/IEEE 15026-2:2022, which specifies minimum requirements for the structure and contents of an assurance case;
- A catalogue of patterns for developing lower-level product-based, process-based, and confidence safety arguments;
- A process for continuously monitoring the status of the safety claims via Safety Performance Indicators (SPIs); and
- A catalogue of templated SPIs, mapped to the claims in the templated argumentation strategy.

In this paper, we only elaborate on the templated top-level argumentation strategy.

The OASCF provides an ecosystem of stakeholders with a roadmap for deploying safe autonomy. The OASCF may be used for safety planning, safety process execution, and safety risk decisions through development, deployment, and sustained system operation. An ecosystem adopts the framework by tailoring it to its risk acceptance criteria[1], deployment goals, autonomy technology, use cases for autonomous vehicles, and regulatory expectations.

The intended audiences for the OASCF are diverse, ranging from machine learning developers to insurance underwriters. These audiences depend on a consistent understanding of residual risk. Many organisations unfortunately juggle disparate views about risk even across their own internal teams. Worse still, regulators responsible for upholding public safety often receive incomplete and optimistic views of risk. The OASCF seeks to bridge these gaps by providing a coherent assessment of residual risk viewed through lenses suitable for different audiences. The template for the structured argument shows how system artefacts generated by different teams support the risk assessment together. Further, the OASCF supports the definition and evaluation of safety-relevant tasks and artefacts throughout the development process, as it includes templates and evaluation criteria for evidence artefacts, as well as the engineering processes used to produce the artefacts.

OASCF harmoniously combines best practices from the following sources into a single compelling argument that is supported by evidence generated by a unified safety engineering process:

- Safety case assessment from ANSI/UL 4600;
- Safety management systems from the FAA (2020);
- Hazard analysis from MIL-STD 882E;
- Engineering rigour from standards relevant for users across industries, such as the Joint Software Systems Safety Engineering Handbook (USDoD 2010)., or ISO 26262;
- Validation of the suitability of autonomy algorithms from ISO 21448 (SOTIF);
- Monitoring of safety performance indicators from UL 4600; and
- Lifecycle safety processes best practices.

[1] Risk acceptance criteria is defined in ISO 21448 as a criterion representing the absence of an unreasonable level of risk. Acceptance criteria may be quantitative or qualitative. One example of such criterion given in the standard is the maximum number of incidents per hour. What is considered ‘unreasonable’ depends on the organisation’s risk appetite.

4.2 Positive Trust Balance

Autonomous vehicles that rely on machine learning for life-critical functions like tracking pedestrians and surrounding traffic hold the promise of being safer than human drivers. To this end, Positive Risk Balance (PRB) shall be demonstrated. Demonstrating the achievement of PRB before the deployment of autonomous vehicles without sufficient lagging metric data to provide high confidence in an acceptable safety outcome is difficult. The argument employed by the OASCF is inspired by (Koopman and Wagner 2020) and supports a responsible deployment decision despite the lack of lagging metric data. This approach relies on practical evidence collection, supporting the expectation of sufficiently low risk rather than the unrealistic requirement of conclusive proof that a risk target has been met right from the outset. Positive Trust Balance (PTB) proposes using a combination of validation, engineering rigour, post-deployment feedback, and safety culture to predict deployment risk confidently. Further, responsible deployment requires monitoring key assumptions and safety metrics post deployment and to continuously calibrate risk.

The top-level claim of our templated high-level structured argumentation strategy is "*The autonomous driver is safe enough to operate in the considered operational design domain*". An autonomous driver is a system embedded in a vehicle, which navigates and operates the vehicle without human intervention. An autonomous driver uses various sensors, such as cameras, radars, and lidars, to perceive the operating environment of a vehicle, make decisions, and perform driving tasks without human input. Autonomous drivers can have different levels of automation, ranging from Level 0 (no automation) to Level 5 (full automation). Implementing the PTB approach, the OASCF relies on three pillars of argumentation (see Figure 4). The first pillar argues about having confidence in the safety argument given the employed processes in the organisation. The second pillar argues that the system is safe by design, namely that the system was engineered and tested using rigorous processes while considering the identified hazards. The third pillar argues that the hazards uncovered during system operation are appropriately addressed.

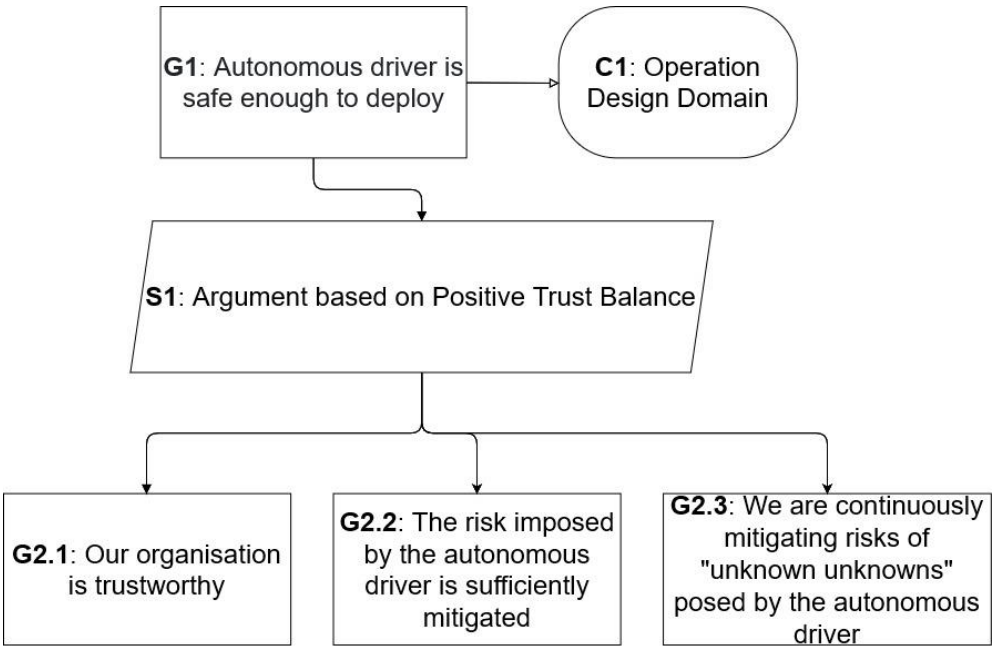

Figure 4 ~ Top-level Claims of OASCF in GSN

4.3 Live It Right

OASCF requires a robust and transparent safety culture to ensure sound technical outcomes and build public trust over time. The "Live It Right" pillar (see Figure 5) argues that the organisations building and operating the AV are safe. This means that these organisations have strong safety management and safety culture and that they deploy novel technology responsibly. Establishing this allows the reader of the safety case to trust the rest of the argument.

This section of our framework covers critical issues such as:

- Avoiding setting unreasonably low initial quality and validation goals based on an argument that post-deployment updates will fix bugs. Such an argument is not aligned with the Positive Trust Balance approach. Instead, risk acceptance criteria must be aligned with industry best practices, regulatory requirements, and societal expectations. Our framework uses guidance from the German Ethics Commission (BMVI 2017) on setting nuanced quantitative risk acceptance criteria and safety integrity level approaches from safety standards.
- Using the lack of maturity in accepted practices in some areas (e.g. still evolving best practices for safe machine learning) as an excuse for not following well-known best practices for more traditional aspects of the system, such as functional safety.

Risk acceptance criteria, organisational principles, and safety management plans form the basis of a safety management system (SMS) safety policy and safety promotion processes. SPIs tied to "Live It Right" claims track deviations from these policies and processes.

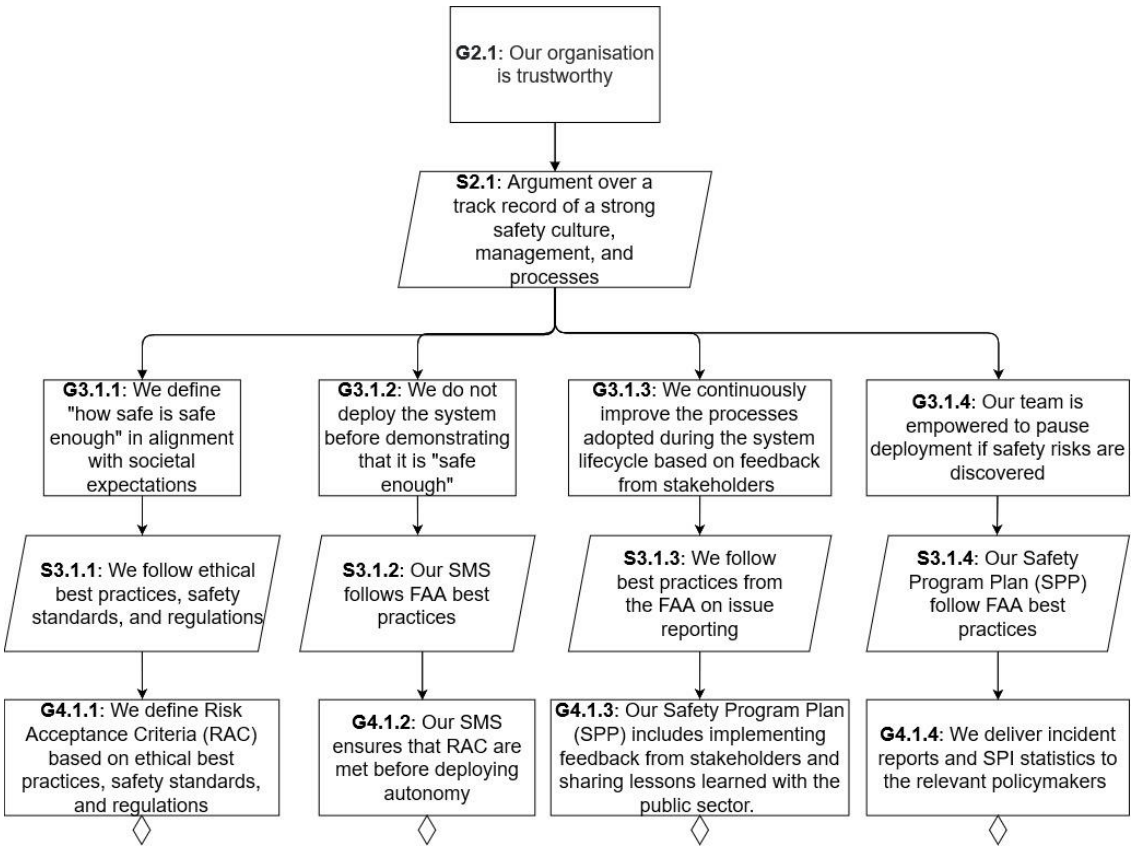

Figure 5 ~ The "Live It Right" Argumentation Pillar

4.4 Engineer It Right

This part of the safety case (see Figure 6) argues that the autonomy technology is safe by design, and the development and safety assurance of the system is grounded on rigorous engineering development practices. The OASCF demands that the manufacturer of the autonomous system builds their technology in accordance with industry standards for identifying and mitigating safety risks. Defence standards, like MIL-STD-882E, refer to this as a system safety process, whereas the FAA refers to it as safety risk management. These standards define how to:

1. Identify loss events that can occur within a defined concept of operations;
2. Assess and track hazards that could lead to these loss events and understand how the autonomous vehicle can cause these hazards can occur;
3. Define safety-critical functions that prevent hazards from occurring;
4. Implement safety-critical functions with a level of rigour appropriate to the severity of risk being mitigated;
5. Verify that implementations are correct, following the guidance from standards; and
6. Validate that hazards have been mitigated as intended in operation, following the guidance from standards.

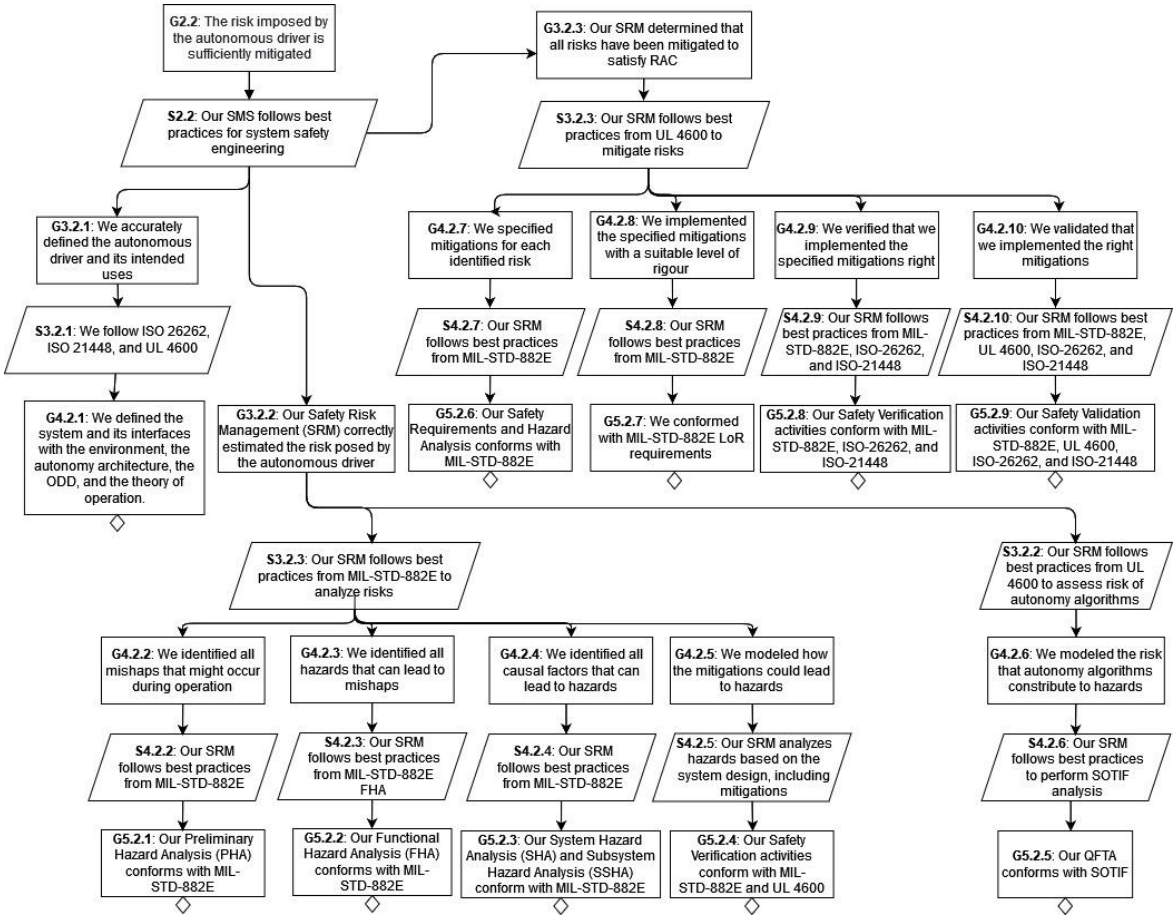

Figure 6 ~ The "Engineer It Right" Argumentation Pillar

Safety standards provide tables specifying methods for the above tasks and the work products they generate. The OASCF relies on all these work products and independent assessments of the manufacturer's conformance with relevant standards. Our framework

provides templates for each of these work products. However, while we publish the OASCF, we do not intend to publish the templates.

Among others, quantitative fault tree analysis (QFTA) generates the evidence supporting this argumentation pillar. For each identified hazard, the safety case refers to a quantitative fault tree (QFT) that models our best understanding of how causal factors can lead to a hazard, along with expectations for how often these causal factors will occur in practice. The QFT consists of a function (typically Boolean) describing how the combination of conditions referred to as causal factors can lead to a hazard.

The structure of a QFT is generated by well-known hazard analysis techniques such as those used in MIL-STD-882E. First, a preliminary hazard analysis (PHA) defines what loss events can occur and lists what hazards can lead to these loss events. Second, the functional hazard analysis (FHA) allocates the occurrence of hazards to functions in the autonomy stack and builds the fault tree. Third, the system/subsystem hazard analysis process allocates functions to components in the architecture, and can employ techniques such as failure modes, effects, and criticality analysis (FMECA) to define further what causal factors need to be tracked, including hardware failures and software defects.

Risk targets or estimates can be allocated to each node in the fault tree. Initially, this can be used to define a risk budget for the autonomy stack and operations. However, risk can be estimated — and the risk budget refined — with feedback from SPI monitoring during simulation and on-road testing. During initial deployment, statistical deviations between SPIs seen in validation and those seen in real-world operations can serve as a trigger for an issue for the safety management system to resolve. Further, we use the QFTs to define safety requirements on the operations of an autonomous vehicle in a way that can be monitored using telemetry, examined for anomalies and gaps in hazard analysis, and used to improve the safety case over time.

Mitigating hazards will also require instituting operational controls. Reasons for this include the need for machine learning in safety-critical functions, the open-ended nature of requirements on autonomous behaviours, and unavoidable uncertainty in our models of the environments in which autonomy operates over time. An understanding of all such situations is the aim of methods in MIL-STD-882E and ISO 21448. For example, in automotive, operational control might include limitations on driving routes that stay away from school zones, bike lanes, and densely populated urban areas. Other operational controls require preventative maintenance and diagnostics to ensure that safety-critical functions in the autonomous vehicle are dependable. While testing autonomous vehicles, operational controls usually include safety drivers with dependable takeover mechanisms outside the scope of the autonomous vehicle itself, for example, as defined in SAE J3016.

4.5 Operate It Right

“Operate It Right” is the argument that the operation of the autonomous driver is safe (see Figure 7). The claim that an autonomous system is “operated right” is based on two assertions: that all the operational controls are correctly in place, and that the QFTA accurately describes the rates of causal factors that occur in practice.

Together, these two claims build an argument that the safety assurance processes of an SMS is effective. An SMS uses safety assurance processes to evaluate the continued effectiveness of mitigations in the field, and to identify previously unknown risks and hazards.

Process audits can provide evidence that operational controls are properly in place. However, audits measure inputs to the operating system, not the system's outputs.

Consequently, OASCF also demands evidence from safety performance indicators, or SPIs, that monitor the behaviour of the autonomous driver in operation.

The data from the operation are recorded, analysed via SPIs, and acted upon by the manufacturer to assure safety in the field. SPIs are central to the FAA's safety assurance strategy, and are also used by UL 4600 to monitor safety case assumptions in real-world operation. The analysis of SPIs supports the following goals:

- Identifying areas where safety performance can be improved;
- Evaluating the effectiveness of operational safety risk controls;
- Increasing confidence in the system to validate the expected risk of operation; and
- Minimising harm by identifying unknown issues and risks before losses occur.

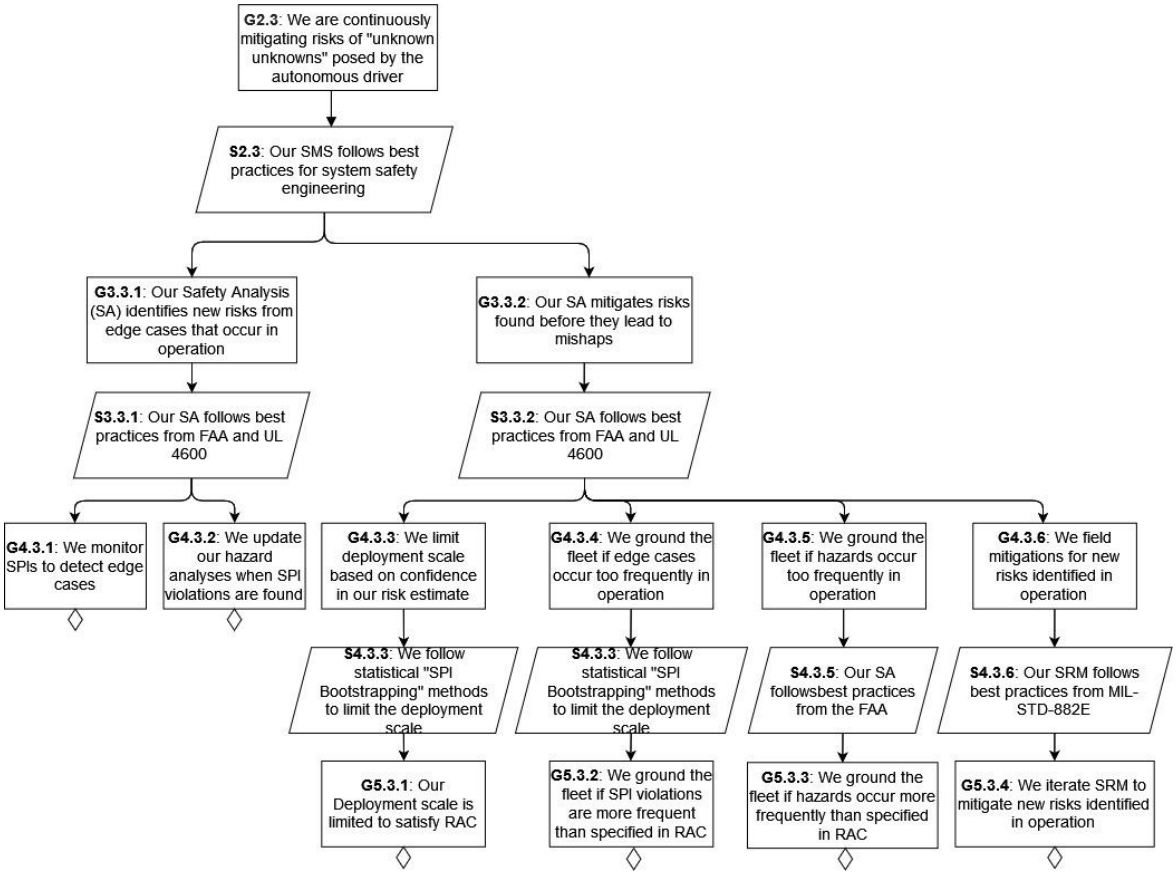

Figure 7 ~ The "Operate It Right" Argumentation Pillar

5 Related Work

5.1 Safety Cases for Autonomous Systems

Several works in the literature have discussed how to structure safety arguments for AVs.

Guarro et al. (2017) propose a risk-informed safety case framework for unmanned aircraft systems, accompanied by a validation and verification framework. The safety case framework connects the risk-based claims with the evidence generated by applying the validation and verification (V&V) framework. This work focuses on integrating V&V

results into the engineering-based argument but does not address broader safety argument concerns such as safety culture and post-deployment monitoring.

Hawkins et al. (2021) propose a process for the Assurance of Machine Learning in Autonomous Systems (AMLAS), accompanied by a set of GSN-based safety argument patterns. However, their proposed patterns argue that an ML-based component satisfies its allocated safety requirements. These safety argument patterns may be used to refine the claims in OASCF.

Wishart et al. (2023) propose a safety case framework for AVs, which argues that an AV is safe for its intended implementation based on three argumentation pillars. Like OASCF, their framework considers UL 4600, ISO 26262, ISO 21448, and different AVSC reports. The first two pillars — "SMS" and "Design Methods" — reassemble our "Engineer It Right" pillar, as they argue that an SMS and a safety engineering process based on standard requirements and best practices are in place during system development. The third pillar of the SCF proposed by Wishart et al. (2023) is "Scenario-Based Testing", arguing about a scenario-based testing process employed to evaluate the performance of the AV. This third pillar can be used to develop OASCF's Goal G5.2.8: "*Our Safety Verification activities conform with MIL-STD-882E*" further.

5.2 Safety Case Frameworks for Autonomous Vehicles

As shown in the following, our OASCF aligns with existing SCFs released by AV companies. Our OASCF can be seen as a *meta* SCF, which can be adapted by AV companies based on their specific needs, namely specific audience, use cases, used standards, and norms. Also, an adapted SCF shall be in line with the company's safety process. Once the OASCF is enhanced with company-specific information, it can be used to create safety cases for different systems.

The first organisation to use a safety case for self-driving cars was Uber ATG. Uber ATG published its safety case framework after an independent review by a team at ECR. This framework has since been adopted by Aurora Innovation, who acquired Uber ATG in 2021. The initiative taken by Uber ATG and Aurora has hastened the adoption of safety cases throughout the self-driving industry, prompting safety case development at Locomation, TuSimple, Waymo, Zenzic, and other organisations. Aurora (2023) published the first four levels of their high-level structured safety argument in a GSN format. The argument has five argumentation pillars, which can be mapped to our three pillars.

Aurora's, "*The self-driving vehicle is proficient during nominal operation*", "*The self-driving vehicle is fail-safe*" and "*The self-driving vehicle is resilient to reasonably foreseeable misuse and unavoidable events*" high-level claims map to the "Build It Right" pillar in OASCF, whereas their "*The self-driving vehicle and safety processes are continuously evaluated and improved*" maps to our proposed "Operate It Right" argumentation pillar. Our "Live It Right" argumentation pillar is reflected in Aurora's "*The enterprise is trustworthy*" claim.

The safety case framework proposed by Waymo has the absence of unreasonable risk as its top-level goal (Favaro et al. 2023). Waymo considers three types of hazards, namely architectural, behavioural, and operational. To assess the residual risk associated with the identified risks, risk acceptance criteria are defined. Similar to what OASCF proposes, the safety approach proposed by Waymo is dynamic and relies on safety as an emergent development property, safety as an acceptable prediction and observation, and safety as continuous confidence growth. This means that, similarly to OASCF, Waymo's

framework requires that rigorous engineering practices are used during system development, actual and simulated performance is combined to measure safety performance indicators, and feedback loops exist to address safety issues occurring after deployment and to confirm the residual risk computed before deployment. The framework proposed a three-layered argumentation:

- Performant, secure, and robust hardware platform (drive-by-wire);
- Safe and responsible driving behaviour (autonomy); and
- Safe deployment and operations (operations).

OASCF's "Engineer It Right" argumentation pillar may be used to develop the first two layers, whereas the OASCF's "Operate It Right" pillar is one way to structure Waymo's third argumentation layer. To evaluate the evidence and the arguments within the safety cases generated while using Waymo's Safety Case Framework, Waymo proposed a Case Credibility Assessment. OASCF instead proposes templates and checklists, for generating and evaluating credible arguments and evidence.

The Zenzic safety case framework (Zenzic 2021) supports the development of safety cases for AV testing, explicitly targeting the CAM Testbed in the United Kingdom. The framework provides its users with guidance on how to follow standards required for being able to test in the UK, such as ISO 26262, ISO 21448, PAS 1881:2020, PAS 1883:2020, while explicitly avoiding defining specific methods or tasks, instead providing useful guidance around the organisation and content of the safety case. While the Zenzic framework does not provide any concrete structured argument specified in a graphical notation, it proposes three argumentation strategies. One strategy argues about the functional safety and the safety of the intended functionality of the system under test. OASCF's "Engineer It Right" argumentation pillar may be used to construct a structured argument for this pillar.

The second argumentation strategy proposed by Zenzic framework concerns operational safety. It aims to demonstrate that a test vehicle can operate safely within the defined environment, relying on evidence that considers the interaction of the test vehicle with the operating environment, including the route, safety driver or operator, passengers and other road users. Here, OASCF's "Operate It Right" argumentation pillar may be used to construct a convincing, structured argument.

Zenzic's third argumentation strategy considers security threats, such as physical access or via electronic and telecommunications means (cybersecurity). Whereas OASCF does not yet consider security assurance, we plan to extend it with an argumentation strategy concerning cybersecurity in the future, as mandated by ISO/SAE 21434:2021. The Zenzic framework also advocates for safety cases to be live documents, and to be updated when previously-unknown hazards are uncovered. However, it does not propose a solution for how to specify a live safety case. Whereas OASCF is based on PTB, the Zenzic framework argues that the identified residual risk is As Low As Reasonably Practicable (ALARP)[2].

[2] A term from UK safety legislation.

6 Discussion

The need for specifying safety case interfaces. The autonomous driver is usually embedded in the vehicle, usually receiving information from the sensors or sending commands to the actuators in the vehicle. This means that the safety case created for the autonomous driver, based on our proposed OASCF, shall be embedded in the safety case of the vehicle. Consequently, to ease the integration between safety cases, safety case interfaces shall be defined, specifying the assumptions and guarantees of a safety case related to, for example, the followed regulations and standards, the considered failure models, the achieved SIL/acceptance criteria and validation targets, the failure modes, or the safety requirements/safety goals/performance requirements.

The challenge of implementing "live" safety cases. The level of rigour needed for implementing live safety cases will take time to be achieved by autonomous driving development companies. Specifying SPIs and implementing an automated CIA are time-consuming and challenging tasks. This is why ECR has a catalogue of SPIs that can be customized for different projects, a toolchain, and a process supporting automated CIA. Further, to close/approve a live safety case before deployment, relevant stakeholders would also need to approve the defined SPIs and the implemented automated CIA. Another challenge is to specify a process for responding in case SPIs are invalidated while pondering between system availability and system safety.

7 Summary and Future Work

In this paper, we introduced OASCF — a safety case framework for autonomous systems, based on the PTB rationale. OASCF has been developed based on ECR's vast experience gained while working with different stakeholders in the AV industry, from developers to regulators. Safety concepts and argument templates for autonomous vehicles quickly rely on implementation and organisation-specific details. A significant difficulty in generating generic argument templates is finding a balance between providing argument content and over-prescribing the details through templating. The templated structured argumentation within OASCF is high-level, focusing on processes, allowing program-specific details to be incorporated through references and evidence artefacts. Whereas the first nine levels from the OASCF we published in this paper showcase how to use various standards in order to build up a strong safety argument, the argument patterns accompanying the templated argument, which support safety engineers in further-developing templated arguments, showcase how to reference all work products mandated by the standards. Next, we plan to publish some of the argument patterns. In this paper, we discussed how our proposed OASCF covers the claims of existing SCFs, such as the ones proposed by Aurora, Uber, and Waymo. Our OASCF can be seen as a *meta* SCF, which different companies can instantiate.

Soon, we will apply OASCF for use cases and system types. Also, we plan on developing the SCF further so that it addresses European and American regulatory requirements. For example, European Regulations (EU) 2019/20144 and (EU) 2022/1426 pose explicit performance requirements for AVs/Automated Driving Systems (ADS). (EU) 2022/1426 also guides the derivation of scenarios relevant to the Operational Design Domain (ODD) of the ADS. In the United States, AV companies must satisfy licensing requirements, comply with specific safety standards, and follow testing protocols. Whereas our SCF guides its users towards compliance with a set of safety standards, we plan to enhance it to

facilitate compliance with cybersecurity standards, such as SAE J3061_201601 and ISO/SAE 21434. Further, we will publish a catalogue of state-of-the-art safety argument patterns supporting the creation of arguments supporting our high-level templated argument. Also, another next step is to propose a catalogue of SPIs that, together with our OASCF can be used to specify live safety cases.

Our proposed SCF is *open*, meaning that everyone is invited to:

1. use our SCF to create their safety cases, and
2. contribute to perfecting the SCF so that all AV stakeholders can benefit from the safe deployment of AVs.

Hereby, we invite everyone to find assurance defeaters attacking our safety claims, argumentation, strategies, and referenced evidence and send them to us to enhance the OASCF.

Correspondence Address

Michael Wagner: mwagner@ecr.ai

Carmen Carlan: ccarlan@ecr.ai

References

Aurora. (2023). *Aurora's Safety Case Framework*. Aurora Innovation Inc. 2023. https://safetycaseframework.aurora.tech/gsn. Accessed 3rd February 2024.

BMVI. (2017). *Ethics Commission Report: Automated and Connected Driving*, Bundesministerium für Verkehr und digitale Infrastruktur, the Federal German Ministry of Transport and Digital Infrastructure. https://perma.cc/6UBX-KH5G. Accessed 3rd February 2024.

Cârlan C., Gauerhof L., Gallina B., and Burton S. (2022). *Automating Safety Argument Change Impact Analysis for Machine Learning Components*. In :IEEE 27th Pacific Rim International Symposium on Dependable Computing (PRDC) (pp. 43-53). IEEE.

(EU) 2019/20144. (2019). *Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users*. European Union.

(EU) 2022/1426. (2022). *Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 laying down rules for the application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles*. European Union.

FAA. (2020). *Safety Management System*. FAA Order 8000.369C, June 2020. US Department of Transportation Federal Aviation Administration.

Favaro F., Fraade-Blanar L., Schnelle S., Victor T., Peña M., Engstrom J., Scanlon J., Kusano K., and Smith D. (2023). *Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk*. Waymo LLC. arXiv preprint arXiv:2306.01917. https://arxiv.org/ftp/arxiv/papers/2306/2306.01917.pdf. Accessed 3rd February 2024.

Guarro S., Yau M. K., Ozguner U., Aldemir T., Kurt A., Hejase M., and Knudson M. (2017). *Risk Informed Safety Case Framework for Unmanned Aircraft System Flight Software Certification*. In :AIAA Information Systems-AIAA Infotech @ Aerospace. https://doi.org/10.2514/6.2017-0910.

Hawkins R., Paterson C., Picardi C., Jia Y., Calinescu R., and Habli I. (2021). *Guidance on the assurance of machine learning in autonomous systems (AMLAS*). Assuring Autonomy International Programme (AAIP), University of York, UK. arXiv preprint arXiv:2102.01564. https://arxiv.org/pdf/2102.01564.pdf. Accessed 3rd February 2024.

ISO 21448. (2022). *Road vehicles — Safety of the intended functionality*. ISO 21448:2022, 1st Edition, 2022. International Organization for Standards, Geneva.

ISO 26262. (2018). *Road vehicles – Functional safety*. ISO 26262:2018, in 12 parts, 2nd Edition, 2018. International Organization for Standards, Geneva.

ISO/IEC/IEEE 15026-2. (2022). *Systems and software engineering — Systems and software assurance — Part 2: Assurance Case*. ISO/IEC/IEEE 15026-2:2022, 2nd Edition, 2022. International Organization for Standards, International Electrotechnical Commission, Geneva, and Institute of Electrical and Electronics Engineers, Piscataway, NJ.

ISO/SAE 21434. (2021). *Road vehicles — Cybersecurity engineering*. ISO/SAE 21434:2021, 1st Edition, 2021. International Organization for Standards, Geneva, SAE International, Pittsburgh.

Koopman P., Ferrell U., Fratrik F., and Wagner M. (2019). *A safety standard approach for fully autonomous vehicles*. In: Computer Safety, Reliability, and Security: SAFECOMP Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Proceedings 38 (pp. 326-332). Springer International Publishing, New York.

Koopman P. and Wagner M. (2020). *Positive trust balance for self-driving car deployment*. In: Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops: DECSoS, DepDevOps, USDAI, and WAISE, Proceedings 39 (pp. 351-357). Springer International Publishing, New York.

MIL-STD-882E. (2023). *Department of Defense Standard Practice — System Safety*. MIL-STD-882E w/Change 1, 27 September 2023. US Department of Defense.

MISRA. (2019). *Guidelines for Automotive Safety Arguments*. Motor Industry Software Reliability Association (MISRA), Norwich.

PAS 1881. (2020). *Assuring the safety of automated vehicle trials and testing — Specification*. PAS 1881:2020, 1st Edition, 2020. The British Standards Institution. https://www.bsigroup.com/globalassets/documents/pas/pas1881_final-design-proof.pdf. Accessed 3rd February 2024.

PAS 1883. (2020). *Operational design domain (ODD) taxonomy for ADS specification*. PAS 1883:2020, 1st Edition, 2020. The British Standards Institution.

SAE J3016. (2023). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. J3016, Work in Progress, 2023. SAE International, Pittsburgh.

SAE J3061_201601. (2016). *Cybersecurity Guidebook for Cyber-Physical Vehicle Systems*. J3061_201601, 2nd Edition, 2016. SAE International, Pittsburgh.

UL 4600. (2023). *Evaluation of Autonomous Products*. ANSI/UL4600, 3rd Edition. Underwriters' Laboratories, Northbrook, IL.

USDoD. (2010). *Joint Software Systems Safety Engineering Handbook*. Version 1.0, August 27, 2010. US Department of Defense.

Wishart J., Zhao J., Woodard B., O'Malley G., Guo H., Rahimi S., and Swaminathan S. (2023). *A Proposed Safety Case Framework for Automated Vehicle Safety Evaluation*. IEEE International Automated Vehicle Validation Conference (IAVVC), Austin, TX. https://www.researchgate.net/publication/374842966_A_Proposed_Safety_Case_Framework_for_Automated_Vehicle_Safety_Evaluation. Accessed 3rd February 2024.

Zenzic. (2021). *Safety Case Framework: The Guidance Edition*. Zenzic-UK Ltd. https://zenzic.io/innovation/safety-and-security/safety-case-framework/. Accessed 3rd February 2024.

ISSN 2754-1118 (Online)
ISSN 2753-6599 (Print)

Human Factors in Functional Safety Assessment

Assessment of Human Interactions and Behaviour

Jonathan Wiggins

1981 Consultants, Trowbridge, Wiltshire. UK.

Abstract

This paper provides a discussion on the physical and emotional factors which humans bring into the design, execution, and maintenance of a Functionally Safe System. It aims to bring these elements together in a structured manner and suggests potential assessment criteria for use by practitioners in this field to allow the assessment of Human Factors into an IEC 61508 or related safety system, though no limitation is intended as the basic methods may be modified for other fields. This paper does not go into specific applications, rather looks at a generic approach as a basis from which sector specific factors may be added. This paper provides a basis for the building of Key Performance Indicators to baseline, measure, and compare differing elements in the Human Factor assessment against a suggested target for a given Systematic Capability level. This is in the context of an audit-based assessment attempting to minimise the effect of bias on the process to bring the often-subjective elements into an objective and measurable format.

1 Introduction

Within functional safety there is an increasing understanding that the decisions and processes in the design and operational phases, which are not directly hands-on with the system, have an impact on the overall integrity of the system. This has led to an increasing focus on the techniques and measures used in the design, assessment and operation of systems, and has been crystallised into the concept of Systematic Capability (SC) levels in IEC 61508-2:2010 for hardware and IEC 61508-3:2010 for software. These standards however could still be seen as taking a mechanistic view of human working practices and cultures, and failing to account for systematic failures caused by physical and emotional factors such as stress, fatigue, or poor motivation. In this regard the cybersecurity sphere as defined in standards such as the IEC 62443 series has a greater understanding, as it looks at intentional human actions which cause harm. Whilst it is not implied that all or even most intangible systematic failures will be caused by deliberate actions, the basic approach has merit.

This paper discusses two types of human factor — the tangible and intangible. Tangible factors are those such as ergonomics, information loading and physical layout. In many cases, though, the poor design decision made leading to systematic issues may be caused by Intangible factors such as stress, fatigue, culture, and misunderstanding. Much of the work to date has focussed on the impact tangible factors have on people operating and maintaining

a system, and not the intangible, though a poor appreciation of tangible factors will lead to issues caused by intangible factors in most cases.

Human factors assessments are often subjective in nature as what is clear or acceptable to one person is not to another. This may lead to unclear requirements or outcomes in the functional safety design and assessment process. It is the aim of this paper to account for this in the approach described and suggest the allocation of clear metrics to this field.

To accomplish the metrics-based approach this paper draws on the cybersecurity concept of a Key Performance Indicator (KPI) based framework and set of metrics as described in IEC TR 62443-3-1:2009, which may be taken as a starting point for building an assessment process and asking the correct questions to lead to understanding the human qualities in system design and implementation. It seeks to advise on the manner in which the assessment should take place and the focus of the assessment through the system lifecycle.

This paper, though, is not an exhaustive list of human factor considerations, nor a design guide to implementing Human Factors. Such guides and standards exist, and this should be read in conjunction with them and in the context of the application.

2 Types of Human Factors

Human interaction with a safety system is the most constant aspect in the system lifecycle. From the requirements definition to de-commissioning, human interaction exists. The nature of this interaction can fall into two categories, and hence types of factors to be considered in a system lifecycle:

- Tangible Factors (Section 8).
- Intangible Factors (Section 9).

Tangible Factors describes the humans' interaction with the physical system. This includes the design, ergonomics, information loading and work loading on the operator(s) created by the system and the ability of the operator to both understand what the system is telling them and to react in the correct manner to the information.

In the early stages of a system lifecycle this could also include the workstations being used by, for example, the engineers, analysts, and safety practitioners to create the designs and requirements for the system and interoperate the outputs of models and simulations being run as at the stage pre-realisation; this virtual interaction is a critical part of the system.

The Intangible Factors describe the interaction of humans with other humans, the culture, and the environment under which work is being undertaken. This factor is present throughout the system lifecycle but is especially prominent in the early phases and during phases where abstract work (such as theoretical analysis) is undertaken.

3 Effect of Lifecycle

Through the lifecycle there is a shift in the human factors between the intangible to tangible. This can be seen as being due to the different types and focus of the interaction with the system. For an example, at the concept and initial design stages the system is, by and large hypothetical ideas and concepts. Therefore, the intangible factors play a greater role as a

stress on the mind at this stage could lead to a critical error being introduced at an early stage. It is suggested that, at the beginning of a lifecycle, the concept phase is 10% tangible and 90% intangible factors; with the 10% representing factors such as workstation design and office layout. By operation there is a far greater focus on the physical interaction with an actual system. This shifts the focus to a suggested 80% tangible factors, such as the layout of the control stations, and 20% intangible for the culture and demands of the organisation operating the systems. The shift from intangible to tangible comes at system realisation. Errors, though, are often only detected at the point at which the system is operated for the first time. To counter this, an equal focus is placed on tangible and intangible factors at realisation.

As the system ages, human understanding of the system decreases with memories fading and the natural turnover of staff within an organisation creating an increased reliance on written records. This is seen as causing a rise in intangible factors in the latter stages of the system lifecycle. Modifications to the system, especially in the latter stages of the lifecycle, are therefore to be considered in a similar way to the initial concept and design phases in so much as the intangible consideration become higher as the process which leads to the modification is as important as the effect of the modification on the overall result. This is because there is often a high degree of conceptual re-engineering or reverse engineering to be done in addition to the modifications. Decommissioning is a specific case where the shift to focus on intangible factors occurs late in the lifecycle. Being the point farthest from the design phase, the system may have passed through multiple parties prior to decommissioning. This increases the probability that incorrect or convenient assumptions will be made in the decommissioning process, leading to additional risk of harm.

Table 1suggests an apportioning of tangible and intangible factors against the lifecycle stage as defined in IEC 61508-1:2010, Figure 2 “Overall Safety Lifecycle”.

Table 1 ~ Suggested Tangible vs Intangible Focus Through IEC 61508 Lifecycle

IEC 61508-1:2010 Stage		Tangible	Intangible
1	Concept	10	90
4	Overall Safety Requirements	20	80
10	Realisation	40	60
12	Overall Installation and Commissioning	50	50
13	Overall Safety Validation	50	50
14	Overall Operation Maintenance and repair	80	20
15	Overall modification and retrofit	50	50
16	Decommissioning or disposal	40	60

4 Information Flow

In understanding human factors, it is important to understand the flow of information passing through the human function being assessed. The first level is the tangible flow of information from the system to the person and from the person to the system, but there is also the flow of information between people to bring into account. A simplified framework

is illustrated in Table 2, which suggests a flow of information from supplier to customer and from a supervisor/manager to a subordinate. This should be framed in terms of both a project or operational environment and an organisational environment to understand where potential conflicts and gaps exist. Table 2 defines the roles as singular, but this may refer to a team of people or multiple people. In many cases the Object Owner changes through the lifecycle with the supplier and customer being the previous or next in the chain. This is overseen by Process managers who often oversee the larger picture of a project or operation of a system. It is important these interactions are understood.

Table 2 ~ Simplified Interpersonal Information Flow Scenarios

Name	**Object Supplier**	**Process Manager**	**Object Owner**	**Object Customer**
Role	Defines starting points, prerequisites and input targets into system for Object Owner to work on. May also act as source of knowledge of Object Owner	Allocates and oversees the work by the object owner and coordinates activities with Object supplier and Object Customer.	Actions the required work within the defined process points by the object supplier and Process Manager.	Receives the completed work package from the object owner along with output points, prerequisites and targets
Interaction	Supplies information, and input targets to Object owner	Oversight, internal target setting	Specialist knowledge transfer with Owner or Supplier.	Progresses the information to the next stage in the process. May become the next Object owner.

At each stage and level described in the flow of information there is a point of handover and interaction. These points are often set by:

- Time Constraints (imposed or required, e.g. shift changes),
- Geographic constraints,
- Knowledge constraints,
- Business demarcation.

The full and complete handover of decisions and knowledge at each stage is a critical step to be monitored. That this is not done under conditions where parties are under pressure due to time, workload, or interpersonal conflicts should be considered. The Process Manager should be aware of this and looking out for potential points of stress in the inter-personal system as well as withheld information. The assessment of human factors should pay particular attention to these points in the process and the transfer point as a key indicator of the health of a system.

Examples of regular transfer points may be a shift change or the phase move from one part of the lifecycle to another (department changes). Changes in Process Management should also be looked at as a shift.

At the higher level, the Object Supplier and Customer may be separate stakeholders within the process. The flow of information between, for example, the team implementing the safety system and the end users is often through many layers of people. These interpersonal information flows are another source of stress which should be paid attention when assessing the quality of the Human Factors.

5 Nature of the Human Factors Assessment

Translating human factors requirements and considerations into measurable and meaningful outcomes is achieved through an assessment process. With human factors there may be a high degree of subjectivity in the results. This precludes a mechanistic assessment in many cases as a ranking done by a single practitioner, especially if there is a relationship between the assessor and the assessed, may reflect the bias and opinion of that person. This is not a deliberate action, rather a case of unconscious bias. The assessment of human factors therefore should be done by a group of assessors, where each may be from a different background (business role, profession, culture, etc.) better to give a diverse view of the inputs given. The group nature of the assessment lends itself to an audit process to determine the effectiveness of the approach being taken.

Auditors should be knowledgeable in the subjects, and as independent as required from the process being audited. IEC 61508-1:2010, 8.2.17 Tables 4 and 5 set out recommended levels of independence for functional safety assessment. These represent a good framework, although it is suggested that the factors for determination of X1 and X2 as per IEC 61508-1:2010 8.2.16 are considered in terms of:

- Greater degree of organisational complexity;
- Lack of experience in similar field;
- Greater degree of organisational novelty; and
- Greater degree of design novelty.

It is important to be mindful that many of the subjects being audited are sensitive and therefore great care should be taken to not breach confidentiality or cause undue stress, which could bias the audit results. There are counter-arguments as to the effect of independence on the results. These are that, whilst the familiarity with the people can relieve the stress of discussing sensitive subjects with strangers at the greater degrees of independence; the fact that they *are* strangers balances this by lessening the likelihood that the auditee feels that the information disclosed will leaked. It is considered that these factors broadly balance overall, and have a lesser impact that may at first be foreseen.

It is suggested that, in addition to compliance to IEC 61508-1:2010 Clause 8, audits are conducted in accordance with ISO 19011:2018, or a similar recognised format. Two auditors and ideally at least one observer should be present, with the observer there to arbitrate and ensure a fair process. Audits should be conducted ideally on an individual basis, unless there is a need for a group discussion. This should be determined in advance and communicated. All information must be kept in confidence and results anonymised to prevent the results being traced to individuals. The latter encourages honest views to be expressed over the tendency to say what is expected by the organisation.

6 KPI Target Levels

6.1 KPIs

Using KPIs is a structured approach to the targeting and assessment of information within the context of performance parameters. The following is derived from IEC TR 62443-3-1:2009, with modifications to the wording to suit the safety domain. These are further drawn from ISO 22400-1:2014. This is felt an expedient means, as these are both industry recognised and transferrable between safety and security domains should the need arise.

KPIs should be structured in a manner which is:

- Measurable;
- Repeatable;
- Specific to the assessment context;
- Contain achievable and specific outcomes;
- Quantifiable / Comparable; and
- Meaningful to all parties.

The process for building performance metrics starts with the understanding of the purpose of the assessment. This allows the analysis of what each requirement is asking in context with the safety objectives structure and other related requirements.

To analyse the standard requirements and derive metrics for each of them, the following steps are defined:

- Identify the Steps;
- Evaluate the context;
- Identify the actions and owners; and
- Identify follow up actions required.

KPIs are derived from performance metrics based on assessment context and objectives, or elements' structure. This process is well described and documented in multiple sources. ISO 22400-1:2014 is an industry recognised industrial method for building KPIs; it is essential that a recognised method is utilised in building KPIs.

Some performance metrics alone have enough weight and priority that they are worthy to track individually as a KPI. In some other cases, tracking a group of related performance metrics as one KPI may be more meaningful. Whatever the case, this approach allows for the use of either method. After completing the performance metrics for each of the safety requirements or objectives, the project team should then proceed to build KPIs. These indicators will be created based on either an individual performance metric or a group of them as per the identified risk reduction objectives. The process and criteria for each KPI should be recorded to assure consistency and repeatability.

The KPIs should be set in the context of the lifecycle phase under consideration, and actioned items be achievable within the same phase. Where a KPI crosses phases, separate measures across the different phases should be established and actioned accordingly. It is best that each phase can close out all outstanding actions. Carry-over actions can confuse the end of phase transfer, as these could be either due to cross phase KPIs or the failure to

achieve a KPI. By keeping each KPI and actions within a set period this is a clearer process with clear ownership.

Once established, KPIs should be regularly reviewed for achievement, relevance, and quality of both the KPI and the outcomes. Where needed, changes are made to KPIs so as to improve the safety outcomes.

6.2 KPI Target Levels

When setting KPI performance targets one difficulty is creating a uniform assessment criterion. To do this within the context of functional safety a simple 1 to 5 scale may be adopted. This allows the targets to be integrated with other factors in standard 5x5 risk tables.

The suggested targeting levels are as per Table 3:

Table 3 ~ Human Factors Consideration Levels

Level	Description
L1	Limited consideration. Top level only with minimal risk reduction requirements.
L2	Detailed consideration. Target should aim to mitigate the top 25% of risks to an acceptable level
L3	Detailed Consideration. Targets should aim to mitigate the top 50% of risks to an acceptable level.
L4	High consideration. Targets should aim to mitigate the top 75% of risks to an acceptable level
L5	Very High consideration. Targets should aim to mitigate the top 90% of risks to an acceptable level.

7 Using a Structured Approach to Human Factor Design

The use of a structured approach to Human Factors design is critical to identifying key points and ensuring these are acted upon. This should be referenced to the questions:

- Who will be using the system?
- How will the system be used?
- When will the system be used?

The use of appropriate standards and guidelines is strongly advised at all levels. Standards such as ISO 9241-210:2010, IEC 62366-1:2015+A1:2020, and EN 614-2:2000+A1:2008 describe this subject and the necessary precautions to be taken. When human factors is being considered in both the design and assessment it is critical that this is in a structured manner and integrated into the technical design to prevent elements being missed or causing conflicts down the line which may be resolved in an ad-hoc manner.

When being assessed, the use of standard methods and a structured integrated approach should be credited for both tangible and intangible factors, as this ensures that human stress and the room for induced errors is minimised.

8 Tangible Human Factors

8.1 Preface

Tangible and intangible human factors have been described briefly up to this point in the overall context of a functionally safe system. In the next two sections it is intended to flesh these out in more detail with key focus areas. These are generic and consideration should be given towards the nature of the role being undertaken, the working environment and mobility of the operator need to be accounted for in the correct detail. The following areas are though headings to be used as starting points to determine the detailed factors.

8.2 Clarity

The attributed clarity of a system describes the clarity with which functional information and controls are presented to the operator. This may in terms of both the system process steps and the operator cognitive steps.

Examples are a control panel lid laid out in a form of a process flow, or a start and stop button being co-located and clearly labelled or marked with an infographic. In the design phase this may also involve tools being used guiding the practitioner through the steps of the process.

Considerations for clarity can include:

- Logical layout of relevant information.
- Appropriateness of information being displayed considering both the task being undertaken and the location of the display relative to the task inputs.
- Ambiguity of information display to the operator. This may be the use of unclear language or symbols. Examples include where the operators first language is not the one the system works in.
- The size and readability of data. This may include size, font contrast and colours used. Considerations including undiagnosed visual impairments such as mild colour blindness should be made here.

8.3 Ergonomics

The ergonomics describes the attribute od control location, expected function and work loading on a human operator. The Health and Safety Executive (HSE 2013) in the United Kingdom describes this as:

> *Taking account of ergonomics and human factors can reduce the likelihood of an accident. For example, in the design of control panels, consider:*
>
> - *The location of switches and buttons — switches that could be accidentally knocked on or off might start the wrong sequence of events that could lead to an accident.*

- *Expectations of signals and controls — most people interpret green to indicate a safe condition. If a green light is used to indicate a 'warning or dangerous state' it may be ignored or overlooked.*
- *Information overload — if a worker is given too much information, they may become confused, make mistakes, or panic. In hazardous industries, incorrect decisions or mistaken actions have had catastrophic results.*

8.4 Interaction

Interaction describes level of required operator input. This may be for control inputs or for inspection inputs, e.g. monitoring CCTV.

Assessment starts with the level of input required. Too much input may lead to operator overload and a step being missed. Too little or repetitive input may lead to inattentiveness and the operator not noticing a vital warning requiring attention.

It is often desirable to have multiple operators in a safety context. This leads to additional interactions between the operators and the operators and the system. Both should be considered. When using multiple people, each role needs to be clearly defined with understood overlaps if these exist. This prevents one leaving the role to the other, or both attempting the same task.

8.5 Cognitive Step

Cognitive step attributes describe the amount of mental processing needed to take a command and execute the correct next step. This could also include a requirement for specialist training and knowledge of the system which must be learned. A clear warning and controls may be provided but, if the response requires many years of training and an easily misjudged response, a high degree of cognitive step is required to respond.

Each additional cognitive step potentially leads to errors due to human judgement, learning or emotional response. It is therefore desirable that the number of cognitive steps to implement a safety action should be at a minimum. Factors in assessing cognitive step could include:

- Intuitive nature of controls;
- Level of human judgement needed to maintain control;
- Number of different methods to achieve the same aim; and
- Level of process understanding needed.

In many security systems a level of cognitive step is used as a part of the authentication process. This requirement should be considered along with the needs of safety in any assessment.

8.6 Tangible Human Factors — Level Targeting

To be able to specify proportionate and targeted KPIs for tangible human factors, it is important to have an easy to use and understand system. To this end, this document proposes a "TH Level" linked back to the desired SIL[1] level of risk reduction through

[1] Safety Integrity Level

systematic capability. The use of the SC levels is described in IEC 61508-2:2010 and IEC 61508-3:2010 as the means to reduce systematic errors arising in the design operation and maintenance of the system.

Bringing out the top-level targets, it is important to determine the target for each factor under consideration. This is to ensure that each is assessed both individually and as a part of the overall assessment. Suggested TH Levels are as Table 4.

Table 4 ~ Tangible Human Factors Target Levels (TH-L)

Metric	SC1	SC2	SC3	SC4
Clarity	TH L2	TH L3	TH L4	TH L5
Ergonomics	TH L2	TH L3	TH L4	TH L5
Interaction	TH L2	TH L3	TH L4	TH L5
Cognitive Step	TH L3	TH L4	TH L5	TH L5
Structured Approach	TH L2	TH L3	TH L4	TH L5

8.7 Tangible Human Factors — Level Assessment

When reviewing the KPIs and performance, a simple method of assessing the achieved TH level is important. This should be assessed alongside the target mitigation level and target SC level, thus as the severity of the consequence rises so the required outcome is higher.

For each outcome the maximum level achievable is that set as L. Where a non-compliance is found this can be used to deduct the L level as suggested in Table 5. This may be linked to a 5x5 risk matrix through Table 5.

Table 5 ~ Tangible Human Factors Compliance Levels

Assessed Compliance						
None	**5**	No L	No L	No L	No L	No L
Low (<50%)	**4**	L-3	No L	No L	No L	No L
Moderate (50–75%)	**3**	L-2	L-2	L-2	L-3	L-3
High (76-98%)	**2**	L-1	L-1	L-1	L-2	L-3
Very High (>98%)	**1**	L	L	L	L-1	L-1
		1	**2**	**3**	**4**	**5**
		Severity				

An example is that the TH-L KPI is targeted to a TH-L4 in order to achieve SC3 for Ergonomics. Achievement KPI is assessed as being High however the severity of the gaps identified is moderate (3), this means that the Actual achieved level which can be claimed is TH-L3 for ergonomics.

9 Intangible Human Factors

9.1 Preface

The intangible factors describe the interaction of humans with other humans, the culture of the place and the environment of the place. This is often a subjective matter though indicators of issues are available for observation.

9.2 Stress

(HSE 2023) defines stress as: "*the adverse reaction people have to excessive pressures or other types of demand placed on them*". Workplace stress may be a result of stress accumulated both in the workplace and outside the workplace. It is important to recognise stress in the workplace and take this into account. The HSE further describes this:

> *Workers feel stress when they can't cope with pressures and other issues. Employers should match demands to workers' skills and knowledge. For example, workers can get stressed if they feel they don't have the skills or time to meet tight deadlines. Providing planning, training and support can reduce pressure and bring stress levels down.*
> *Stress affects people differently — what stresses one person may not affect another. Factors like skills and experience, age or disability may all affect whether a worker can cope.*

9.3 Environment

Environment describes the area in which the worker is conducting the activities relating to functional safety which impact the ability of the worker to conduct the duties required. There are two aspects to the environment.

First is the physical. Is the physical space suitable in terms of:

- Space;
- Temperature;
- Noise;
- Light;
- Access to welfare (toilets, washbasins, etc);
- Safety of environment; and
- Tangible human factors surrounding the workstation?

The second is the intangible factors of pressure and practice which can lead to:

- Bullying;
- Discrimination; and
- Demeaning.

9.4 Organisational Culture

According to (HSE 2019). Culture can be best understood as "*the way we do things around here*". Culture forms the context within which people judge the appropriateness of their

behaviour. An organisation's culture will influence human behaviour and human performance at work. Poor safety culture has contributed to many major incidents and personal injuries.

The culture of an organisation may be looked at in terms of

- Buy-in and action on safety at all levels.
- Allowance for diversity of opinion.
- Decisions being made on a rational or evidence basis.
- Whistleblowing of poor practice and actions being taken.
- A proactive and measured approach to implementation of safety.

9.5 Misunderstanding

Misunderstanding seeks to describe the level to which the workers at all levels are appropriately trained and informed on the work they are undertaking. The training element encompasses:

- Formal learning;
- Vocational learning; and
- Practical and theoretical understanding of the functional safety principles.

For learning to be effective, it is essential that information on the system application, requirements, and other working areas is distributed adequately. Working with a lack of information and communication could lead to inconsistent decisions being made between teams and a misunderstanding of what the aims of the system are. Practices such as:

- Silo or Over-the-wall working;
- Inter-team communication both verbally and written;
- Regular and well-mannered communications;
- Stockpiling of information;
- Systems to allow collaboration being used, but not over-relied on;
- Linguistic barriers;
- Cultural barriers; and
- Distance, including remote working.

9.6 Policy Implementation

Many, if not all, organisations will have policies covering these areas. From the point of view of safety, it is critical to understand that a policy is a statement of intent. The key question to be asked is, “Is this being enacted?”. When looking at intangible human factors, the reality must be assessed, and not just the intention of a policy or procedure.

The presence of a policy is a good sign, and must not be discounted in the overall assessment. It is recommended, though, that the assessment assumes in the first instance that there is no policy and no process. This leaves a set of questions:

- Is the leadership proactively setting the correct tone and goals?
- Are the workers working responsibly and appropriately?
- Is the workplace suitable for the task being undertaken?
- Are all personnel adequately informed and educated?

- Are the goals realistic considering the above?

Evidence for evidence-based policy and decision making may be found in reference to implementation of standards such as ISO 27500:2016, ISO 26000:2010 and, in a more limited capacity, ISO/IEC38500.

9.7 Performance Targeting

As with tangible human factors, intangible factors may be targeted into 5 levels based on the required systematic capability (SC). For Intangible factors these are the IH levels shown in Table 6. KPIs should be set to target the required IH level in each area, and aim to achieve total compliance. There is in these areas, though, a strong desire to drive improvement. KPIs which have longer term objectives may be set, but these must not lose sight of the basic functional safety requirements. Suggested IH Levels are as Table 6.

Table 6 ~ Intangible Human Factors Target Levels (IH-L)

Metric	**SC1**	**SC2**	**SC3**	**SC4**
Stress	IH-L2	IH-L3	IH-L4	IH-L5
Environment	IH-L2	IH-L3	IH-L4	IH-L5
Culture	IH-L2	IH-L3	IH-L4	IH-L5
Misunderstanding	IH-L3	IH-L4	IH-L5	IH-L5
Policy	IH-L3	IH-L4	IH-L5	IH-L5

9.8 Intangible Human Factors — Level Assessment

Assessment of compliance for IH Levels is different from the TH levels. This is because it is very difficult to assesses the impact severity of each factor in isolation as, for example, the culture may drive stress and misunderstanding. To this end, the frequency that a particular factor is seen versus the KPI level should be assessed. Table 7 provides suggested guidance on this and the means to convert to a 5x5 risk assessment.

Table 7 ~ Intangible Human Factors Compliance Levels

Assessed Compliance						
None	**5**	No L	No L	No L	No L	No L
Low (<50%)	**4**	No L	L-3	L-3	L-3	L-2
Moderate (50–75%)	**3**	L-3	L-3	L-2	L-1	L
High (76-98%)	**2**	L-2	L-1	L-1	L	L
Very High (>98%)	**1**	L-1	L	L	L	L
		1	**2**	**3**	**4**	**5**
		50%	**75%**	**85%**	**92%**	**98%**
		Frequency				

An example is that the TH-L KPIs are targeted to an IH-L3 to achieve SC2 for Stress. Achievement KPIs is assessed as being to a Moderate degree in all cases, High degree in 9.5 out of 10 cases, and Very High degree in 9 out of 10 cases. Therefore, the assessed level is IH-L3.

Where goal setting KPIs are used, this method may equally be used to assess progress and the current position in relation to the desired end goal. The assessment, though, must be against the current position and current safety targets, and not consider any future developments as these may or may not happen.

10 Conclusion

The impact of people in functional safety requires greater attention than it currently receives if systematic capability metrics are to improve. This document draws attention to the potential areas of concern and proposes a method to assess the impact of human factors and behaviour but does not propose that this is the definitive method rather a framework into which a definitive method and metrics can be derived.

The results of the process described within this document could be seen as an addition to IEC61508-2:2010, Annex B or IEC 65108-3:2010, Annex A as an additional technique and measure to ensure Functional Safety.

Initially, it will be for individual practitioners to assess the impact based on the application being assessed, fill in this framework, and set appropriate KPIs and targets. Longer term though, further work is needed involving experts in functional safety, human factors, and human behaviour. This work needs to involve original equipment manufacturers, system integrators, end users, and policy makers across industry sectors to provide a rounded and consistent approach to this issue and integration into the Functional Safety process.

Correspondence Address

Jonathan (Jon) Wiggins may be contacted via jon.w.1981@gmail.com.

Acknowledgments

I wish to acknowledge Audrey Canning, Ed Lambert and Phil Williams for their invaluable assistance and encouragement in reading, checking and commenting on this paper.

I wish also to thank Phil Williams for the suggestion that this become a formal paper.

References

EN 614-2. (2008). *Safety of machinery — Ergonomic design principles — Part 2: Interactions between the design of machinery and work tasks*. EN 614-2:2000+A1:2008, 1st Edition, 2000, as amended 2008. CEN, the European Committee for Standardization, Brussels.

HSE. (2013). *Ergonomics and human factors at work: A brief guide*. UK Health and Safety Executive. https://www.hse.gov.uk/pubns/indg90.pdf. Accessed 26th January 2024.

HSE. (2019). *Organisational culture: Why is organisational culture important?* UK Health and Safety Executive. https://www.hse.gov.uk/humanfactors/topics/culture.htm. Accessed 16th January 2024.

HSE. (2023). *Work-related stress and how to manage it*. UK Health and Safety Executive. https://www.hse.gov.uk/stress/overview.htm. Accessed 16th January 2024.

IEC 61508-1. (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems — Part 1: General requirements*. IEC 61508-1:2010, 2nd Edition, 2010. International Electrotechnical Commission, Geneva.

IEC 61508-2. (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems — Part 2: Requirements for electrical/electronic/programmable electronic safety-related systems*. IEC 61508-2:2010, 2nd Edition, 2010. International Electrotechnical Commission, Geneva.

IEC 61508-3. (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems — Part 3: Software requirements*. IEC 61508-3:2010, 2nd Edition, 2010. International Electrotechnical Commission, Geneva.

IEC 62366-1. (2020). *Medical devices — Part 1: Application of usability engineering to medical devices.* IEC 62366-1:2015+A1:2020, 1st Edition, 2015, as amended 2020. International Electrotechnical Commission, Geneva.

IEC TR 62443-3-1. (2009). *Industrial communication networks — Network and system security — Part 3-1: Security technologies for industrial automation and control systems*. IEC TR 62443-3-1:2009, 1st Edition, 2009. International Electrotechnical Commission, Geneva.

ISO 9241-210. (2019). *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. ISO 9241-210:2019, 2nd Edition, 2019. International Organization for Standards, Geneva.

ISO 19011. (2018). *Guidelines for auditing management systems*. ISO 19011:2018, 3rd Edition, 2018. International Organization for Standards, Geneva.

ISO 22400-1. (2014). *Automation systems and integration — Key performance indicators (KPIs) for manufacturing operations management — Part 1: Overview, concepts and terminology*. ISO 22400-1:2014, 1st Edition, 2014. International Organization for Standards, Geneva.

ISO 26000. (2010). *Guidance on social responsibility*. ISO 26000:2010, 1st Edition, 2010. International Organization for Standards, Geneva.

ISO 27500. (2016). *The human-centred organization — Rationale and general principles*. ISO 27500:2016, 1st Edition, 2016. International Organization for Standards, Geneva.

ISO/IEC 38500. (2015). *Information technology — Governance of IT for the organization*. ISO/IEC38500:2015, 2nd Edition, 2015. International Organization for Standards, International Electrotechnical Commission, Geneva.

This collation page left blank intentionally.

ISSN 2754-1118 (Online)
ISSN 2753-6599 (Print)

About the Safety-Critical Systems eJournal

Purpose and Scope

This is the Journal of the Safety-Critical Systems Club CIC (SCSC), ISSN 2754-1118 (Online), ISSN 2753-6599 (Print). Its mission is to publish high-quality, peer-reviewed articles on the subject of systems safety.

When we talk of systems, we mean not only the platforms, but also the people and their procedures that make up the whole. Systems Safety addresses those systems, their components, and the services they are used to provide. This is not a narrow view of system safety, our scope is wide and also includes safety-related topics such as resilience, security, public health and environmental impact.

Background

When the Safety-Critical Systems Club (SCSC) was set up over thirty years ago, its objectives were to raise awareness of safety issues and to facilitate safety technology transfer. To achieve these objectives, the club organised events, such as Seminars and an annual Symposium, and published a newsletter, Safety Systems, three times a year.

The Newsletter has, in addition to news, opinion, correspondence, book reviews, and the like, also carried articles discussing current and emerging practices and standards. The length of such articles is limited to about two and a half thousand words, which does not allow an in-depth treatment. It was therefore decided to add a third string to our bow and supplement the events and newsletter with a journal containing longer papers. The journal is now published here, as the Safety-Critical Systems eJournal, and comprises at least two issues a year.

Content Sources

Sources include the outputs of SCSC working groups; solicited technical articles and topic reviews; submitted articles on new analysis techniques, discussion of standards, and industrial practice; and guidelines and lessons learned. If you wish to contribute, please see, "Information for Authors".

Types of paper include, but are not limited to:

Technical Articles: Written by practitioners and describing practical safety assurance techniques and their industrial applications.

Integration Studies: Written by practitioners reporting upon successful (or otherwise) synergies achieved in practice with other assurance domains, e.g. systems engineering, reliability/availability/maintainability engineering, resilience, human factors, security, and environment.

Position Papers: Written by, or on behalf of, Regulators, Standardisation Organisations, or other official bodies, setting out their position on a topic, e.g. the interpretation of a particular standard or regulation.

Review Articles: Papers highlighting recent developments and trends in some aspect of safety-critical systems or of their use in a particular industrial sector.

Historical Articles: Papers describing the development of safety assurance in an industrial sector; how we got to where we are today.

ISSN 2754-1118 (Online)
ISSN 2753-6599 (Print)

Perspectives: The authors' personal opinions on a subject, e.g. whether to use statistical methods in particular scenarios.

Reports: The lessons learned from incidents or the outcomes of trials with a description of scenarios, or methods, and a discussion of the results obtained.

Working Group Outputs: Written by Safety-Critical Systems Club Working Groups to include discussions, underpinning theory, or guidelines.

Copyright and Disclaimer

Letters to the Editor

The editorial to the first issue of this journal said, "*You may find some of this material controversial, or you may think that it does not go far enough. Subsequent issues of this journal will have provision for readers' letters to the Editor responding to individual papers.*" Such a letter should be no more than 1000 words in length (not counting title, attribution, or references). That would take up no more than two pages of the journal. Note that a letter should ideally address a single concern with few, if any, external references.