Evidencing responsibility for use of AI in safety critical systems

Philippa Ryan[1], Zoe Porter[1], Laura Fearnley[1], John McDermid[1], Ibrahim Habli[1], Joanna Al-Qaddoumi[2], Phillip Morgan[2]

[1]Centre for Assuring Autonomy, University of York
[2]York Law School, University of York

Abstract

In recent years there has been much dialogue surrounding concepts of "responsible AI" in areas such as ethics, fairness and risk of existential harm from generative AI. However, this dialogue is rarely targeted at AI-based Safety Critical Systems (AI-SCS), which have many unique regulatory and disciplinary challenges compared to other domains. Safety engineers are being increasingly required through regulation to evidence responsible use of AI, but the discipline lacks the conceptual clarity or methodology to do so.

This inter-disciplinary paper uses philosophical models of responsibility (moral, causal, role and legal) to provide clarity for the discipline of safety engineering. We consider AI-SCS relevant challenges, including causal responsibility gaps, the risk of a human in the loop being unfairly blamed after an AI-SCS accident, and the problem of "many hands" hiding responsibility during development. We propose presenting evidence of responsible AI use via graphical responsibility models, suitable for safety engineers to present as evidence as part of a system safety case. We illustrate the application of our approach with two different contrasting examples. The first is a retrospective accident analysis of the death of a pedestrian in Tempe, Arizona involving an autonomous vehicle. The second is a predictive example for an AI-based clinical decision support tool. We show that by using our approach we can uncover residual risk relating to responsibility shortfalls and improve safety by allocating tasks to the most appropriate responsible actors. We identify complex issues around moral answerability and causal contribution for safety tasks. We conclude that using our models can support safety engineers in demonstrating responsible use of AI.

1 Introduction

AI-Based Safety-Critical Systems (AI-SCS), such as autonomous vehicles, inspection drones, medical diagnosis systems and others, are increasingly being developed and deployed in the real world. AI-based systems provide many technical, ethical and societal challenges for assuring their behaviour, particularly when they include black box Machine Learning (ML) components which obfuscate their functionality. This obfuscation makes them especially challenging to use in a safety critical environment (Ashmore, Calinescu, and Paterson 2021). Recent discourse has emphasised the notion of "responsible AI" (Commission June 2024; Madaio et al. 2024; Baeza-Yates 2024; Kou 2024; Joshi and Morley 2019; Porter et al. 2023), whether for development, initial deployment or in use of AI in general applications. A specific issue for the AI-SCS domain is that concrete evidence is needed to support claims about responsibility (such as by analysis or appeal to process) within a safety assurance case (SAC). This evidence will be scrutinised by a regulator or independent assessor, so it needs to be robust and clearly traceable to a system's safety properties. This means there is a need to both consider what notions of responsibility mean in practice for AI-SCS, and further to demonstrate who is/was responsible and for which safety issues.

In common with most AI systems, AI-SCS have a complex development chain with many actors, including software developers, data scientists, regulators, service providers, and suppliers (Cooper et al. 2022; Porter et al. 2023), whose roles and responsibilities all contribute to safety during design. For many AI-SCS there is limited or no interaction with a human operator once deployed (Sujan 2023; Sujan et al. 2019; Macrae 2022) and there are increased gaps understanding who was responsible and/or liable following an incident or accident (Burton et al. 2020; Lawton et al. 2024). These issues, and others, mean demonstrating responsibility for AI-SCS is a complex and qualitative process.

The notion of responsibility has long been studied in a philosophy and law, usually with a backwards-looking perspective to understand events that have occurred e.g., (Tadros 2018; Coeckelbergh 2016; Ryan Conmy et al. 2023; Thompson 1980). However, it has not been studied in depth in safety engineering, which also requires a forwards-looking perspective to predict potential harm and prevent it or reduce its severity or likelihood. Nor has it been practically applied in the context of AI-SCS. This paper considers how a number of different philosophical senses of responsibility (role, moral, legal and causal (Hart 2008)) apply to the actors involved in developing, assuring, regulating and operating AI-SCS. We build upon previous work (Porter et al. 2023; Lock et al. 2009; Baxter and Sommerville 2011; Stahl, B.C. 2023), but the novelty in this paper lies in considering senses of responsibility in harmony with safety engineering practice and culture. For this we explore how practical graphical representation can help to uncover and clarify responsibility gaps when developing and operating AI-SCS.

Our key research questions are as follows:

- **RQ1** - can responsibility gaps during development and deployment of complex AI-SCS be identified and characterised in a meaningful way for AI-SCS?
- **RQ2** - how can this be presented in a transparent way to support a SAC?

This paper is organised as follows. In section 2 we explore related literature to motivate and ground our approach considering **RQ1**. In section 3 we describe our notation for responsibility models considering **RQ2**. Section 4 presents a real-world example of a fatal collision between a pedestrian and an autonomous car. Section 5 presents an example of an AI-based decision support system in healthcare with a safety issue related to biased data. Finally, we present conclusions in section 6.

2 Related Literature and Clarity for AI-SCS

This section focuses on three areas to address **RQ1**. First we describe safety-critical systems, and the ecosystem in which they are built, regulated and operated. With this in mind, we then consider conceptual notions of responsibility, focusing particularly on their applicability to both safety engineering practice and the use of AI. Finally, we consider known responsibility problems for AI-SCS.

2.1 AI-Based Safety Critical Systems

Safety-critical systems are defined as those whose failure, under certain conditions, can lead to harm to humans or the environment (IEC 2010). AI-SCS include AI components to replace some or all of a human function, such as driverless autonomous vehicles (Favaro et al. 2023; Koopman and Wagner 2017), self-piloted inspection drones (Ryan et al. 2024, 2025), and AI-based medical image diagnostics (Ozturk et al. 2023; Deo et al. 2023) [1]. An AI-SCS is subject to differing legal and regulatory regimes, depending on domain, type, country of use and even degree of autonomy. Rather than attempt to cover all these we briefly summarise some common practices which will broadly apply when developing an AI-SCS.

Typically the manufacturers and operators of an AI-SCS will provide a Safety Assurance Case (SAC) that the system is *acceptably safe* for use in a defined context(Kelly 1998, 2004; Bishop and Bloomfield 1998; Sujan et al. 2016). The SAC contains a series of claims (e.g., "System is acceptably safe") which are decomposed into more detailed sub-claims (e.g., "Software tests are complete and satisfied") supported by specific evidence (e.g., a report detailing the tests and results) (ACWG 2021; U.K. Ministry of Defence 2017). The SAC will be reviewed by a regulator or independent safety assessor who may provide approval based on its content.

To show an AI-SCS is acceptably safe it should be evidenced that system specific risks are at least tolerable (e.g., likelihood and severity are below a certain threshold) and have been reduced as far as possible both in design and operation of the system (U.K. Ministry of Defence 2017). Risk is reduced by following established good engineering practice and safety standards (e.g., (ISO 2019, 2018; RTCA/EUROCAE 2011; U.K. Ministry of Defence 2017; IEC 2010)) to find and resolve bugs, and by design improvements, adding operational controls, or reducing severity of accidents (e.g., sprinkler systems). An important issue is that consensus on good engineering practice is not well established for AI-SCS (Ashmore, Calinescu, and Paterson 2021), despite recent proliferation of standards e.g., over 400 at (Institute 2023), making it particularly difficult to present a justification of responsible development based on agreed convention. A key aim for this paper is how to provide this evidence (**RQ1** and **RQ2**).

A cultural consideration for our work is that good safety practice encourages a *just culture* (Dekker 2018)(U.K. Maritime and Coastguard Agency 2022)(Nævestad, Storesund, and Phillips 2018), in which reporting of incidents or safety concerns is encouraged within an organisation, promoting transparency without fear of blame. In this atmosphere, taking collective and personal responsibility for safety is encouraged, ideally preventing future accidents. We come back to this point in the next section.

Where responsibility is documented, it is typically in a limited sense such as via Responsible/Accountable/Consulted/Informed (RACI) matrices used for project management (Defence 2025). These are used to assign specific high level tasks (e.g., prepare a safety case or undertake safety analysis (Rismani et al. 2023)) to an individual, or describe who needs information (e.g., end user, regulator, Tier 1 supplier). This is not adequate to manage more nuanced and detailed types of responsibility, such as the complex supply chain for ML (see section 2.3), where there is no end user, and if legal liability needs to be established (see section 2.3).

2.2 Clarifying senses of responsibility for AI-SCS

Our conceptual model of responsibility for AI-SCS draws upon a number of different sources which we adapt for this specific domain. This includes Porter and Stahl's formulation for responsibility (Porter et al. 2023)(Stahl, B.C. 2023) and Smith's concept of moral answerability (Smith 2015). The work of Porter et al. (Porter et al. 2023) draws upon Hart's classic taxonomy of the different senses of responsibility (Hart 2008) to identify and clarify different types of responsibility (role, causal, legal and moral), used to describe different aspects relating to AI-SCS development and operation. Role responsibility refers to the obligations an agent takes on in virtue of occupying a role. These obligations can include general moral obligations, such as taking on a duty of care, or specific tasks, such as maintaining patient records. It is typically thought that causal responsibility is another way of referring to causation (Grinfeld et al. 2020)(Sartorio 2007). The standard philosophical view is

[1] The scope of this paper is limited to cyber-physical systems with AI functions designed for specific purpose or task. We do not specifically consider generalised frontier AI models such as ChatGPT (OpenAI 2025), although many of the issues (such as many hands 2.3) may be relevant. We note that frontier models may be used during development, e.g., to generate code or review safety data, but for now we assume they are not used as a primary embedded decision making function in an AI-SCS. Later research may consider this.

that to be causally responsible for some outcome is to be a cause, or to causally contribute, to that outcome.

Legal liability responsibility includes being required to pay financial compensation, or be subject to a legal order, or face punishment (Porter et al. 2022)(P. Morgan. and et. al. 2023). Discourse in moral responsibility often relates it to both accountability (praise and blame) and attributability (where an agent's contribution can be considered voluntary but they are not necessarily worthy of praise or blame) (Porter et al. 2022; Ryan Conmy et al. 2023). Based on our knowledge and experience with safety engineering, just culture (Dekker 2018)(U.K. Maritime and Coastguard Agency 2022)(Nævestad, Storesund, and Phillips 2018), and discussion with practitioners these concepts can be overly condemnatory and restrictive. It is important to note that there is always residual risk associated with operation of AI-SCS. Hence, it may not be considered fair to blame developers and/or manufacturers for an accident if they exercised due diligence in reducing that risk as far as possible. Instead, in this paper we consider moral responsibility to be grounded in answerability, where an agent is morally responsible if they are answerable for what they do. An agent is answerable for something if they are an appropriate target of requests for justification regarding that thing (Smith 2015)(Shoemaker 2011). Importantly, however, being answerable does not necessarily make an agent worthy of blame or praise, these kinds of evaluations depend upon the reasons the agent offers in order to justify their conduct. This is revisited in our examples (Sections 4 and 5).

2.3 Responsibility issues

In this section we summarise literature about a number of responsibility issues which apply to AI-SCS. We do not claim this to be a complete list of issues that will apply, but it represents well known and often discussed issues in the domain.

Responsibility gap The first issue is the so-called "responsibility gap" for AI-SCS (Burton et al. 2020; Matthias 2004; Gunkel 2012; P. Morgan. and et. al. 2023) where it is difficult for developers and manufacturers to be held responsible for behaviour of an AI-SCS which contributes to harm. Responsibility gaps can appear in traditional non-AI domains (Da Silva 2024), but there are two particular features of AI-SCS that compound the problem a) ensuring safe behaviour of black-box AI component, and b) defining precise operating domains (Ashmore, Calinescu, and Paterson 2021). In (Yazdanpanah et al. 2022), the authors argue for the AI-based system itself to be programmed to be responsible, although practical methods to do this are unknown.

Munch (Munch, Mainz, and Bjerring 2023), notes that a responsibility gap could be viewed positively, because it avoids unfairly blaming individuals, and replacing human agency with AI can remove the burden of decision making. We reject the idea of not attempting to clarify and, if possible, reduce responsibility gaps. This is not to punish individuals, but to avoid putting people in the situation where they could face unfair blame, and so that we can learn from mistakes and avoid them in future.

Liability Sink Another related, and contrasting, responsibility issue is when the operator (not developer) of an AI-SCS becomes a liability sink (Lawton et al. 2024). When there are accidents, the operator may absorb blame for the consequences of AI-SCS outputs they weren't responsible for creating, and may not have sufficient understanding to avoid or mitigate their consequences. This also applies when the operator may be required to only make interventions when the AI-SCS cannot, increasing cognitive load (Bainbridge 1983; Weaver and DeLucia 2022).

Lawton et al. (Lawton et al. 2024) describe the situation where a clinician includes the recommendation of a clinical decision support system as part of their decision making for a patient. They may be unduly influenced by the system, or ignore it, and in either case could be held morally answerable and/or legally liable for the output of the AI. In (Porter et al. 2022) it is argued that the AI should provide *information* rather than decisions, and that the patient should be at the heart of decision making. We explore a similar scenario in section 5.

Many hands The "problem of many-hands" was introduced by Thompson in (Thompson 1980) and later studied for conventional high-criticality systems (Thompson 2017; Nissenbaum 1996). To summarise: because many individuals and groups of individuals contribute to decisions, activities and outcomes in complex networks and organisations, it is difficult even in principle to determine who is responsible for them. Thompson notes that some may unfairly avoid blame and others unfairly take blame. Both Cooper (Cooper et al. 2022) and Thompson (Thompson 2017) argue that responsibility should be designed into organisations while systems are developed. Using a role responsibility model, such as the one we propose, is one way this can be achieved for AI-SCS.

The problem of many hands, and the difficulties in identifying responsible agents for machine learning in a non-safety environment is explored in (Cooper et al. 2022)(Cobbe, Veale, and Singh 2023), noting the many different responsible agents. Cooper (Cooper et al. 2022) notes that the inevitability of ML bugs can be used to excuse responsibility for non-safety systems. It is standard safety practice, for traditional software, to perform activities to reduce bugs, such as by static code analysis. In fact most regulatory regimes require it and it should be evidenced as part of a SAC (Hawkins 2013; ACWG 2021; RTCA/EUROCAE 2011). Therefore, we do not think responsibility will or can be avoided by developers of AI-SCS, however it is much more difficult to provide compelling evidence for AI due to its black box nature and complexity.

If we are to be clear about the responsibility of actors during the development of AI-SCS, we need to understand the complex supply, development and operation chain for AI (Edwards 2022; Cooper et al. 2022). During development and operation of an AI-SCS there will be multiple developers, engineers, suppliers of components and data, project managers, operators, investors, regulators etc. and it is infeasible to document and link all decisions they make which could contribute to safety. However, if we model key rela-

tionships and responsibilities with more clarity, we could understand and resolve some issues (such as conflicts, gaps and duplicated responsibilities), prevent safety problems in the future, as well as support accident investigation.

Legal pitfalls and soft law One key aspect is that of legal responsibility and liability following an accident. Safety engineers typically follow "good practice" for developing and operating traditional safety-critical systems, e.g., (RTCA/EUROCAE 2011; ISO 2019, 2018). Not following these could lead to sanction, even if the standards are not legally required, and hence is related to responsibility. One problem for AI-SCS is that there is limited consensus and little evidence of what constitutes good practice for developing safe ML (Ashmore, Calinescu, and Paterson 2021), despite a recent proliferation AI/ML standards (Institute 2023), and an increasing volume of research in this area (e.g.,(Hawkins et al. 2021; Buhl et al. 2024; Ashmore, Calinescu, and Paterson 2021; Favaro et al. 2023)). Therefore, a strong case demonstrating due diligence in reducing risk will be needed.

From a legal perspective, one way of describing the purpose of a SAC is a proactive means of justifying and explicitly detailing the roles and responsibilities of the various individuals involved in a system's lifecycle. This clarifies their duties and outlines their subsequent responsibilities. The SAC works in two ways: both forward-looking and backward-looking. The forward-looking intention is to assess potential risks from the outset and identify the relevant roles responsible for mitigating such harms. However, in the event something goes wrong, the SAC will be used backwards-looking in legal proceedings to assess how and where the risk materialised. In complex negligence-based cases, of specific pertinence to AI-SCS, industry standards and expert evidence may be used to assess the scope of duties, any deviations from the standard, and negligent conduct. Published material or guidance forming the general corpus of knowledge in the field may also be used as a reference and can be the deciding factor in liability claims. Importantly, if a person actively and diligently assesses risks, implements safety procedures, and ensures the relevant guidelines are followed, the person is said to have taken reasonable measures to ensure the safety of their conduct. In the event something undesirable occurs, to the extent that liability is based on fault (and not with strict liability claims), it can be argued it would be unreasonable to find someone who diligently follows good safety practice guidelines (and the law, more generally) liable – so long as they have evidently carried out their responsibility, so far as is reasonably practicable or possible (U.K. General Public Acts 1974).

The recent EU AI Act (Commission June 2024) is intended to close some legal responsibility gaps. However, the Act has been criticised for not considering some of the unique problems of AI, including its dynamic nature, its complex lifecycle, understanding end user rights (Edwards 2022), and not tracing responsibility back to AI developers (Watch 2023). The EU AI Liability Directive (Parliament February 2023) is intended to provide means of pursuing civil liability, but legal review suggests technical issues of identifying causal responsibility (responsibility gaps) are not addressed (P. Morgan. and et. al. 2023).

3 Responsibility Notation

In this section we describe a notation, and give a brief overview of how to develop and analyse models for AI-SCS to address **RQ2**.

Our formulation for representing responsibility is: *Actor(A) is (type) Responsible for Occurrence(O)*. An *Actor (A)* could be an AI-based system, an individual involved in development or operation, or an institution (e.g., company). For AI-SCS it is important to distinguish between human actors and AI-actors as the latter has no agency(Zafar 2024) and cannot be morally answerable, only causally responsible. We also include institutions involved in SCS development as these often have key responsibilities which impact on safety, e.g., regulatory bodies and sub-system suppliers.

Occurrences are characterised as either *Decisions*, *Actions* or *Omissions*. An asterisk (*) is used to indicate where the *Occurrence (O)* is attributed to an AI-based actor. For example, *Automated Driving System (A) is (task)role-responsible for executing the dynamic driving task whilst activated*(O)*.

Our notation elements are shown in Figure 1. These are adapted from existing human task modelling approaches which have been previously for safety engineering (Lock et al. 2009; Baxter and Sommerville 2011) but are limited to physical operational tasks (e.g., human tasks for managing a fire). The novelty in our models is both that they are much more wide ranging, considering AI development chains, regulatory responsibilities, and operation, and also that they show many more types of responsibility for more types of actors. In Figure 1 the three different actor types have different symbols, and an *Occurrence* is represented by a rectangle. Resources, i.e., outputs or outcomes from Actors and Occurrences, are represented using the standard flow chart symbol for documents. One nuance is consideration of whether resource is needed immediately (for example, an AI autonomous car braking system) or at a later unknown date (for example, an AI safety regulation policy). Importantly, there may not be a direct interaction between responsible Actors, other than via a common resource.

The relationships between the elements are as follows:

- **(type) responsibility for** - this specifies variations on the four types of responsibility described in section 2.2
- **uses** - indicates where a resource or occurrence is used by another actor. This can add insight where problems with the resource (which may be due to prior issues undertaking the task to manage it) may have safety implications. We note that although a resource may be *used* it may not necessarily be *required* for an actor to complete their task.
- **subordinate to** - this represents where there is a power relationship or hierarchy between elements
- **association** - this is used where there is a non-specific relationship between elements, for example, a resource associated with an occurrence

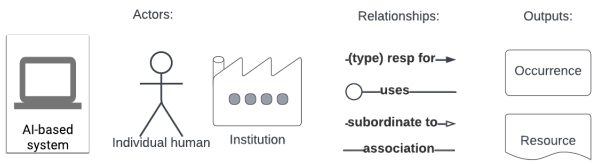

Figure 1: Responsibility model elements

We have introduced the concept of two distinct model types to track responsibility (and responsibility shortfall) throughout the AI-SCS lifecycle. An initial model captures the planned roles, duties, tasks and resources identified for the AI-SCS. This is then used as a template for a series of one or more analysed models, which consider failure modes associated with the occurrences, resources and outputs, thus illuminating where responsibility issues may lie. A simple illustration is in Figure 2 where an AI-based system has the *(task)role* of monitoring whether there is a pedestrian on a potential collision path. If so it creates a collision warning (resource) which is used by the Safety Driver. The safety driver is responsible for intervening to try and prevent a collision when a warning is issued (if necessary).

The initial model is shown in the top line (A) where the actors and their tasks are described. In the bottom row (B), we show the responsibility contributions following a collision. The AI-based system has a causal link only to the insufficient prediction and associated warning (due to lack of agency). It's warning (or lack of warning) conflicts with the safety drivers situational awareness. The safety driver intervenes too late to prevent a collision. They are thus *morally answerable* for their actions even if the AI-based system *causally* contributed to their actions. This highlights a responsibility gap between a causal factor and the driver, as well a potential liability sink 2.3. There may be many different responsibility models, with different potential outcomes, dependent on the issues uncovered. For example, the *Warning of potential collision** in Figure 2 could be *Late* or *Missing*, with different consequences. SCS practice requires the development of forwards-looking analysis models, such as fault trees or hazard logs. These are key pieces of evidence in the SAC, where they are used to demonstrate how potential and credible risks have been identified and how they will be managed (typically looking at physical or functional design). Forwards-looking responsibility models can be used in the same way, this time to identify risks associated actors responsibilities and demonstrate how they will be managed.

We have annotated the tasks and resources with indicative guidewords where a safety shortfall is highlighted. Guidewords are widely used in safety analysis methods e.g., (McDermid et al. 1995; IEC 2016; Leveson N., and Thomas J. 2018; Rismani et al. 2023), to suggest categories of failure. For example, *Too Much* could suggest oversupply of a particular chemical in a processing plant or instead a repeated data signal sent to a car accelerator. A safety engineer will consider the consequences of these failures, which are highly context specific and relate to risk. We have adopted their use here as a familiar tool for safety engineers, regulators, and managers. Having added the failures we then consider any transformations from one type of responsibility to another, e.g. from *task(role)* or *moral(obligation)* responsibility to *causal*. Generally speaking, an individual human or institution could have a responsibility transformation to *(answerability)moral*, *causal* or different types of *liability* where applicable. However, an AI Actor can only become causally responsible as it has no agency. Where there was no shortfall in a responsibility no transformation may be necessary. For example, an individual may carry out their task as expected and there still be an adverse outcome due to an earlier responsibility problem or resource failure. One important point to note is that the transformation in responsibility may or may not considered *fair* to that Actor, particularly if they are a potential liability sink. The models can make such issues transparent so that they can be mitigated if necessary.

In the next sections we extend from a simple model and analysis to include the actor(s) responsible for designing and regulating the AI-SCS, or who provided training data, etc., and build a chain of actors and relationships.

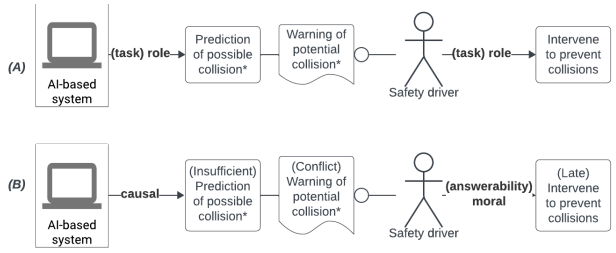

Figure 2: Example of the initial and second analysed responsibility models

4 Example 1 - Fatal collision with an Autonomous Vehicle in Tempe, Arizona

The aim of this section is to show how our models can provide evidence relating to the responsible development and deployment of AI-SCS. We illustrate with a backwards-looking perspective of the fatal collision of an Uber Advanced Technologies Group (ATG) vehicle with an automated driving system (ADS) and a pedestrian pushing a bicycle across a highway in Tempe, Arizona in 2018. The National Transportation Safety Board (NTSB) issued an accident report (National Transportation Safety Board 2019) describing contributory factors from different actors and highlighting safety shortfalls. After the accident Uber ATG described improved safety processes to address some of these (ATG 2018). A sociotechnical analysis of the accident can be found in (Macrae 2022) which also points at many undesirable organisational factors relevant to safety. These reports were the main sources of information for our models.

Uber ATG (an institution) were responsible for developing the ADS, which was an adaptation added to a Sport Utility Vehicle (SUV) from a third-party supplier. The ADS (an AI actor) used a number of ML components to detect and classify objects in the vehicle's path, and also to predict

their trajectories. The accident report states that the classifier failed to consistently categorise the object as a pedestrian with a bicycle. Each time the classification changed, information on the pedestrian's path was lost, and so the ADS could not predict the pedestrian's path correctly to provide a collision warning. The ADS had limited braking capability and relied on driver intervention. The safety driver (human actor) in the vehicle was trained to disengage the ADS in emergency situations and take avoidance action. She was found to be distracted during the accident and intervened too late to prevent the collision. The accident report noted inadequate regulatory control from the National Highway Traffic Safety Administration (NHTSA) over operation of an autonomous vehicle (National Transportation Safety Board 2019).

In 2019, prosecutors said Uber ATG was not criminally liable in the crash. Uber ATG also settled the civil case brought by the victim's family out of court. Uber ATG preferred to use the word "resolved" in place of settled, as they did not admit liability (Fulbrook 2021), although settlements which compromise a dispute often contain provisions that a party does not accept liability. In our opinion, the out of court settlement represents a missed opportunity to test important questions of civil liability and responsibility for incidents involving AI-SCS. Indeed a party may wish to settle a civil claim to prevent a precedent from being set, but also to remove the forensic eye of tort litigation from their operations.

As per section 3 we have two models of role responsibility. The first model shows the actors, and linked tasks (Figure 3), and the second model shows the responsibility findings and shortfalls after the accident (Figure 4).

4.1 First model - example 1

On the left hand side of the initial model (Figure 3) is the regulator NHTSA who have oversight of testing automated vehicles. The production of a safety assessment was voluntary in accordance with the NHTSA regulations, but Uber ATG are still *subordinate to* the regulator as indicated. We recommend that compliance and legal obligations should be included in every initial AI-SCS responsibility model due to their importance in development and operation, and specifically due to grey areas around legal liability (see section 2.3).

We have modelled *Ensuring just safety culture* as a *(moral obligation) role*. This reflects the obligation to embed safety culture within an institution (a concept as well as a set of individual tasks), in accordance with general societal expectations and norms. No individual actor is named as responsible for this, firstly as there are no named individuals in the accident report (potentially a many hands issue), but also we consider it to be a collective obligation for Uber ATG. We note that an individual's degree of responsibility may depend on authority and seniority (Dekker 2018)(Nævestad, Storesund, and Phillips 2018). Other tasks without named individuals *(task) roles* relate to engineering (e.g., *Risk analysis of experimental systems* and *Developing ADS*) and operational occurrences and resources (e.g, *Monitoring safety driver attentiveness*, *Training safety driver*). The ADS has a task of *Warning of collision*. The ADS and safety drivers are shown as separate actors as they have distinct functional roles during operation and are named. The safety driver is sub-ordinate to Uber ATG monitoring, as their procedures allowed for reprimand or other punitive measures if the safety driver was not vigilant to the degree expected.

We have included other road users as a human actor with the qualifier *(1..N)* to indicate that there will be many such individuals. They also bear responsibility in preventing accidents, e.g., by following the rules of the road and monitoring for potential collisions. This does not necessarily indicate a many hands problem, as individuals involved in accidents can typically be identified, although their actions could conflict making it hard to understand causal contribution. We have assumed they are all human actors, but it could be the case that there are other AI-SCS actors with this role.

Although they are named (with no contribution to the severity of outcome) in the accident report, the emergency services are not shown as they do not specifically have a role in *preventing* collisions, which is the scope of our model. Their task instead is to respond *following* an incident, with the potential to reduce the severity of outcome, e.g., by treating those involved. Whilst a model with a broader scope could potentially include these measures, Uber ATG's risk assessment should not rely upon these as good practice would be to *prevent* an incident wherever possible.

4.2 Second model - example 2

Our model has been adapted in Figure 4 to show findings from (National Transportation Safety Board 2019; Macrae 2022; ATG 2018; Wired 2023; Independent 2023). As described in Section 3 we use HAZOP style guidewords to indicate failures. For example, *Warning of collision** becomes *(Late) Warning of collision**. Then we have transformed each of the relationship arrow annotations to reflect the findings, e.g., *(task) role* becoming *causal* or *(answerability) moral* depending on the types of actor involved.

From the prosecution (Independent 2023; Wired 2023; Shepardson 2023) and the accident report, we show that the safety driver's role can be considered as *(answerability) moral* responsible for *(Insufficient) Monitoring for potential collisions* and the late intervention. We have also added the *(criminal) liability* occurrence of *Endangerment* to our model to reflect the outcome of the legal case (Wired 2023; Independent 2023).

For someone to be held morally answerable for an *occurrence*, it is considered a necessary condition for them to both have moral agency and some causal contribution (Porter et al. 2023; Hart 2008; Smith 2015). The ADS has no agency and hence is presented with causal contribution only. This is one area where a potential responsibility gap is highlighted - depending on whether it could be demonstrated that a lack of due diligence from the safety engineers directly contributed to the poor performance. The nature of AI (and in fact any physical system) means there is a chance of random failures (Cooper et al. 2022; Burton et al. 2020). However, principles of risk reduction mean that all reasonable means to reduce the risk should be considered where practicable and possible (Section 2). This is a key area where safety engineers

are encouraged to demonstrate due diligence in their roles and where we believe these models can evidence targeted efforts to reduce responsibility gaps by showing they have maximised risk reduction.

Uber ATG are considered *(answerability) moral* responsible, in that they have agency and it is reasonable for them to justify their actions (as an institution). However, McCrae (Macrae 2022) notes that there was both pressure to deliver (based on fears that the organisation's existence was at stake) and individual staff were disempowered or discouraged from speaking about safety concerns. Hence, individuals might justify that the *(Insufficient) risk assessment* was due to significant pressure on them. The safety issue is that this affects the quality of the ADS, as the risk assessment was not an adequate source of ADS safety requirements to be implemented (the association between the three elements is shown in Figure 4). As noted, Uber ATG are represented as an institution, with no specified individuals. Hence, this is an example of the many hands problem (section 2.3) where we don't have traceability to individual decisions or actions, nor transparency on who would be answerable within the organisation.

One change to our original model is that the specific pedestrian has been included (instead of *1..N* road users), with a causal contribution. The accident report (National Transportation Safety Board 2019) notes that "The pedestrian's unsafe behavior in crossing the street in front of the approaching vehicle at night and at a location without a crosswalk violated Arizona statutes and was possibly due to diminished perception and judgment resulting from drug use". Therefore, if the pedestrian had survived instead they would be held answerable for their actions. The addition of criteria to our analysis method for determining issues such as moral answerability is a potential avenue for further research and may be based on relative importance of causal contributions.

Uber ATG were found not to have monitored the safety driver's awareness, due to underestimation of automation complacency, i.e., where the safety driver trusted the ADS to such an extent that they paid less attention. Also, as noted, there was a complex chain of related occurrences from problems with risk assessment, impacting on the ADS performance, leading to a late warning from the ADS to the driver. This removed a layer of design mitigation that should have reduced the risk of operating the vehicle, and would have reduced the cognitive load on the safety driver. A more prompt warning would have provided mitigation against automation complacency (Bainbridge 1983) and given more time for the safety driver to intervene.

The second model highlights where omissions and shortfalls from a number of responsible actors has increased the burden of risk on the safety driver (thus increasing likelihood of them becoming a liability sink). With greater transparency an operator may be able to better understand the safety responsibility they are taking on (Lawton et al. 2024). We note though, that this example model is based on pre-existing knowledge of an accident. In the next section we explore whether the models can be used identify potential credible responsibility problems.

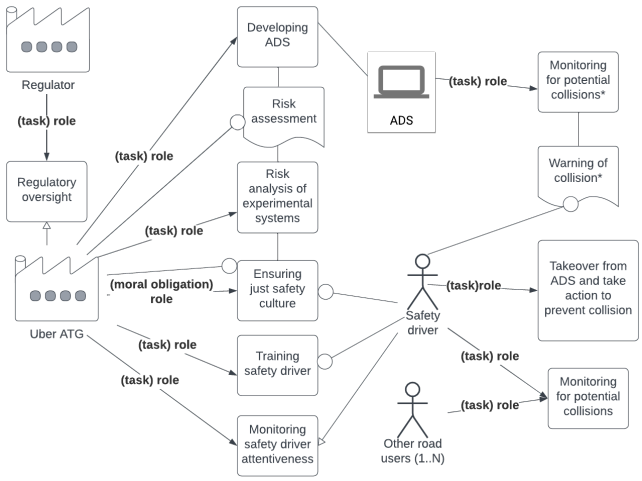

Figure 3: Initial responsibility model showing actors and roles relating to preventing collision

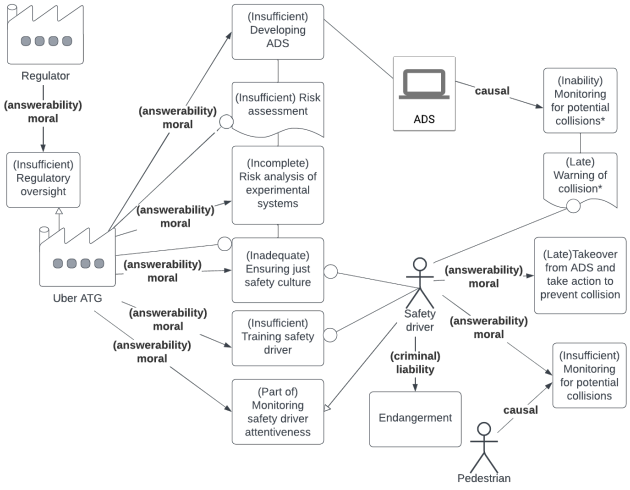

Figure 4: Extended Uber Tempe responsibility model highlighting some responsibility findings from Uber ATG crash in Tempe (National Transportation Safety Board 2019)

5 Example 2 - Diabetes co-morbidity clinical decision support

Our second example is an AI-based diabetes co-morbidity predictor (AI-DCP) being developed to assist clinical diagnoses of patients with Type II diabetes (Ozturk et al. 2023; Ryan Conmy, Ozturk, and Habli 2023; Alonso-Morán et al. 2014). The example demonstrates a forwards-looking responsibility analysis performed during development of an AI-SCS. The aim is to uncover potential safety related failures relating to responsibility gaps, and take mitigating actions where possible. Additionally, we can be transparent about where responsibility gaps remain. As discussed in section 2.1, the SAC should describe and justify any residual risk associated with the system. In a complex AI-SCS supply, development and operating chain with many (potentially) responsible *actors* residual risks will inevitably remain. The forwards-looking analysis will provide evidence for the SAC that steps have been made to identify these in advance of deployment. The analysis can be revised and updated during operation if required.

The AI-DCP predicts a patient's risk of developing a diabetes co-morbidity or having a potentially catastrophic event, such as myocardial infarction (heart attack), within the next six months. It is used during a patient consultation by the clinician, and has been built using an ensemble of ML components (Ozturk et al. 2023) to process recent patient test data and make a prediction of "High" or "Low" risk of a particular co-morbidity. The clinician can choose not to act on predictions made by AI-DCP if they disagree with its assessment or if they have other information, for example, about the patient's preferred treatment, which make alternative courses of action more appropriate (Lawton et al. 2024). The AI-DCP is intended to function as an independent 'second opinion' during consultations, and doesn't specifically recommend treatments. It could influence the decision of the clinician though, either by confirming or contradicting their assessment (Sujan 2023). Therefore, it contributes to the overall safety risk to the patient. A False Negative could lead to no treatment and a False Positive could lead to unecessary medical intervention.

There are larger numbers of missing data values for patients with higher levels of deprivation compared with patients with low deprivation. This could be due to, for example, more difficulties accessing healthcare, or different data gathering procedures. As a consequence this could lead to worse (biased) performance for those patients (Alderman and et al. 2025)(Bender et al. 2021)(Ryan Conmy, Ozturk, and Habli 2023) as aspects of their clinical health are not as well represented in the training data. Management and reduction of bias is guiding principle for the development of AI medical devices in both the U.K. and U.S. (Food and Drug Administration (FDA) and Medicines and Healthcare products Regulatory Agency (MHRA) 2024), and within the EU AI Act (Commission June 2024). Bias can be partially mitigated by careful curation of the training data by the AI developer (Zainuri, Jemain, and Muda 2015; Ozturk et al. 2023), however the impact of some issues cannot be fully understood prior to deployment, nor can we exhaustively test the AI as there isn't a precise specification for what constitutes high and low risk. Ultimately, there could be an inequitable distribution of safety risk. We explore this issue in our analysis, to uncover tasks and responsible actors for both missing training data and resolving it.

5.1 First models - example 2

This section discusses development of the initial models, separated over development (Figure 5) and deployment (Figure 6) for ease of presentation. We have included actors influencing the AI development process, regulators and operators and modelled each of their outputs as occurrences and/or resources.

Some of the modelled human actors (*Clinical Staff*) have the qualifier *(1..N)*. This indicates where there are many actors which we do not have detailed information for, and they are not attached to a specific institution. This is another potential example of the many hands issue (section 2.3). For example, the *Training patient database* (Sohal et al. 2022) is a resource that has been contributed to over many years, by different clinics with different staff, each performing a *Maintaining patient database* occurrence. The *AI software tools* have been developed by many different institutions; for example, commercial tool vendors or open source initiatives. We could extend the model to consider some or all of these individually if, for example, one particular supplier had a more significant role. Multiple institutions would represent many instances of many hands problems.

For regulatory oversight we have listed the health regulator (e.g., Medicines and Healthcare products Regulatory Agency (MHRA) in the UK) and National Institute for Health and Care Excellence (NICE). MHRA provide approval for medical device safety assessments, and NICE provide clinical guidelines which were used to assist in determining the high/low risk factors of different co-morbidities. We have a task for *AI development good practice*, which is also assumed to be assigned to the health regulator. There are, for example, guiding principles developed jointly between the US Food and Drug Administration (FDA), Canada and MHRA (U.S. Food and Drug Administration Oct 2021; Food and Drug Administration (FDA) and Medicines and Healthcare products Regulatory Agency (MHRA) 2024), and regulatory plans from the UK (Transformation Directorate Jan 2025).

In Figure 6 we show the clinician performing a consultation with a patient, using the AI-DCP and electronic and non-electronic patient records for additional information. The patient is included as they have a role during consultation, answering questions about their condition and general health for the clinician. We include that the clinician has a *(moral obligation) role* for *Duty of care*.

5.2 Second model - example 2

In this section we demonstrate how a forwards-looking analysis of the initial models helps develop a revised model with additional tasks (Figure 7) to reduce the safety impact of biased training data. For reasons of space we present a subset of the elements in the figure. The full model(s) can be used

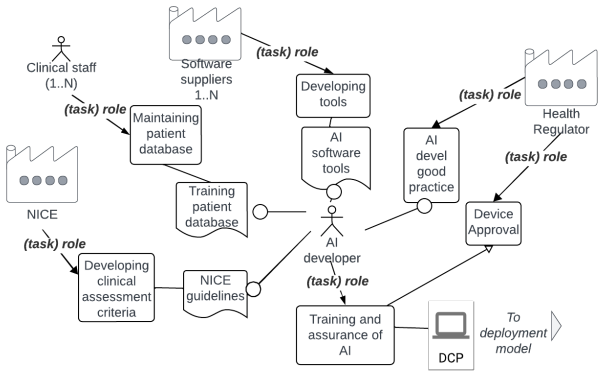

Figure 5: AI-DCP (task)role relationships for development

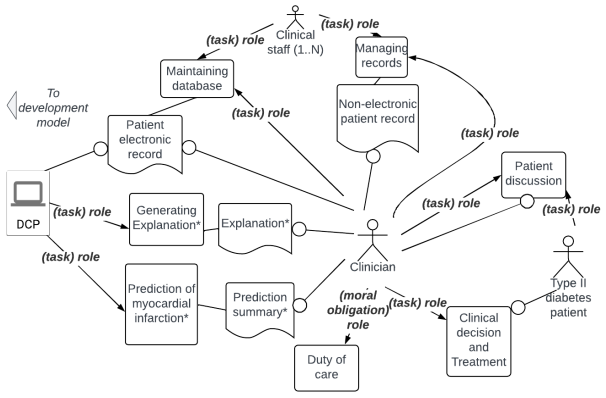

Figure 6: AI-DCP (task)role relationships for operation

as evidence in a SAC that we have considered the safety responsibility issues for this system. We identify credible but hypothetical issues with both occurrences and resources, and think about how they may impact on safety and responsibilities of the actors developing and deploying an AI-SCS. The scope of our analysis is on the impact of problems in the training database, but similar models could be built for, e.g., problems with AI/ML software. As previously discussed, we use guidewords to highlight issues, as well as considering responsibility issues.

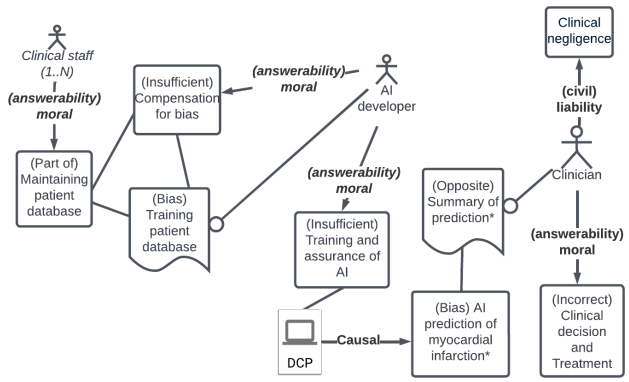

Figure 7: Excerpt from revised model

(Part of) Maintaining patient database implies *(Bias) Training patient database* as data is skewed towards a particular demographic. As noted previously, considering T2D records where the patient demographic was documented we discovered both many more data records from patients with lower levels of deprivation than high (20:1), and, in these cases, a much higher level of missing data ($\sim$15% compared to $\sim$5%). This has a safety and ethical impact as it potentially leads to higher likelihood of incorrect co-morbidity prediction for some patients as their clinical data is poorly represented in training and test sets. The severity of outcome from incorrect treatment remains the same for all (high for myocardial infarction). The exact reasons for the imbalance are unclear, some may be due to different reporting and testing procedures in different clinics (due to funding), and some due to more difficulty accessing healthcare in areas of higher deprivation. In any case, it is good safety practice to reduce a known contribution to risk if possible, even though the distribution of risk across all patients will never be equal due to other individual variability factors (e.g, age, fitness). The model helps identify who might be responsible for risk reduction tasks.

We note that when an individual patient has not had data recorded (e.g., due to missing a test appointment) this has a small impact on the overall data set, meaning it would be hard to point to the actions of a single clinician in failing to secure it as having a significant causal contribution. This is not only an instance of the many hands problem, but is also analogous to the problem of ”insignificant hands” (Hindriks 2022) where one individual makes limited difference, but a large number of them will. A future area to consider is

the proportional relationship between an individuals causal contribution on safety and depth of justification (*moral answerability*) of their actions it is reasonable to ask for. There is a responsibility gap as the clinical staff are not gathering data for the purposes of training AI, rather this task is part of normal patient record keeping. In other words, no one is directly responsible for gathering data for training AI, rather the records have been repurposed for this task. We could still ask for a justification a clinicians data keeping actions if by doing so we uncovered an understanding of behaviour and it led to changes in practice to improve the data. This would be in keeping with the principles of a just safety culture.

Here we propose an additional task in Figure 7 of *Compensation for bias*. This can be reasonably assigned to the AI Developer, as they have the skills to use data imputation (Luo et al. 2022; Ryan Conmy, Ozturk, and Habli 2023) to synthesise additional data to compensate for missing information during training. In Figure 7, we hypothesise that by *Part of* performance of this task they are *(answerability) moral* responsible for *(Insufficient) Training and Assurance of AI*. Importantly, the AI developer may be able to justify this insufficiency if they are still reducing the safety risk from bias as far as possible, but cannot completely eradicate it (Aler Tubella et al. 2023; Alderman and et al. 2025). Thus they have undertaken due diligence in their role by both acting to reduce risk and making transparent the residual risk from bias in the system. This is the essence of risk acceptance supported by a SAC which provides evidence and justification about residual risks, and the measures taken to reduce them, for the AI-SCS. Hence it also supports decision makers who need to accept that risk (e.g., regulators or clinicians).

Following on from this, the bias problem is shown as potentially contributing to *(Opposite) Summary of prediction** and to *(Incorrect) Clinical Decision and Treatment*. In this situation the clinician would be considered *(answerability) moral* responsible for an incorrect diagnosis and treatment plan. In effect, they act as a liability sink, potentially with *(civil) liability*. Using the model they have better information to understand the burden of risk and its causes. Further, they have knowledge to question the validity of AI-DCP outputs for patients with high deprivation, and power to justify their actions if they disagree with the prediction (also considering the patient's preferences).

Establishing a unequivocal causal link between problems with training data and poor predictive performance is not possible due to the black box nature of ML. However, in this section we have illustrated that by introducing a new occurrence/task into our model we can demonstrate that the AI has been developed responsibly to reduce potential bias issues.

6 Conclusions

The key aims of this paper were to present an approach to providing evidence supporting responsible use and development of AI in safety critical systems. We first established a set of responsibility concepts, drawing from philosophical foundations, but considering safety culture and practice (**RQ1**). Then we used this to inform the creation of a graphical notation, presenting responsibilities and relationships between different actors with specific roles in AI-SCS (**RQ2**). We illustrated with two different examples of how it can be used to highlight, and improve, known responsibility issues for AI-SCS, including the problem of many hands and responsibility gaps. An example of bias leading to a safety issue in an AI-DCP was shown, and we illustrated the questions this raises around moral answerability and causal contribution. Here we reflect on the findings and validity of our research in meeting our aims.

We noted that there is a large amount of rapidly evolving regulation for AI, including requirements on demonstrating safe and responsible use of AI. The rapid pace of change may impact on key responsibility issues and tasks which should be included (e.g., for monitoring operational behaviour of AI-SCS, or remote intervention and operation of autonomous shipping (Porter et al. 2023; Kim, Perera, and Sollid 2022)). Much of this regulation also covers wider issues including ethics, transparency and security rather than specifically safety. We have not addressed these directly in our models (although the issue of bias is a cross-cutting ethical concern), but they could be extended and adapted to do so.

Initial feedback from our discussions with experts in the field has been positive, but a formal study with feedback from industrial safety practitioners, or trial on a real system safety case, would be valuable in demonstrating the practicality of the approach. For additional validation we could revisit the healthcare responsibility models (section 5) following an incident or accident. We could test whether the model and analysis had uncovered relevant issues and/or provided valuable evidence relating to responsibility for developers, clinicians and regulators.

In future work we will develop the analysis methodology further, including criteria for, e.g., determining where responsibility could be considered as moral answerability or simply causal. In this paper we have not explored relative causal contribution, for example, including severity and likelihood of safety impact when not performing a role, or other causation theories (Tadros 2018; Bernstein 2017; Lagnado, Gerstenberg, and Zultan 2013), which also relate to answerability and liability (Hitchcock and Knobe 2009). This is a complex area with much disagreement amongst both philosophers and safety practitioners.

Finally, our work is potentially generalisable to other SCS or frontier AI, but our focus was on cyber-physical AI-SCS due to the known issues around responsibility and immediate regulatory concern amongst safety practitioners.

7 Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (EP/W011239/1) and the Centre for Assuring Autonomy, a partnership between Lloyd's Register Foundation and the University of York. We would like to thank Paul Noordhof (Philosophy Department, University of York), Tom Lawton (Bradford Teaching Hospitals) and Berk Ozturk (AAIP, University of York) for their invaluable insight and comments.

References

ACWG. 2021. Goal Structuring Notation Community Standard. Technical Report SCSC-141C v3.0, Safety Critical Systems Club.

Alderman, J. E.; and et al. 2025. "Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations". *The Lancet Digital Health*, 7: 64–88.

Aler Tubella, A.; Coelho Mollo, D.; Dahlgren Lindström, A.; Devinney, H.; Dignum, V.; Ericson, P.; Jonsson, A.; Kampik, T.; Lenaerts, T.; Mendez, J.; and Nieves, J. 2023. ACROCPoLis: A Descriptive Framework for Making Sense of Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 1014–1025. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Alonso-Morán, E.; Orueta, J.; Esteban, J.; Axpe, J.; González, M. l.; Toro, N.; Loiola, P.; Sonia, G.; and Nuño-Solinís, R. 2014. The prevalence of diabetes-related complications and multimorbidity in the population with type 2 diabetes mellitus in the Basque Country. *BMC public health*, 14: 1059.

Ashmore, R.; Calinescu, R.; and Paterson, C. 2021. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, 54(5): 1–39.

ATG, U. 2018. Safety Report Supplement Internal and External Safety Reviews. Online: accessed November 2023.

Baeza-Yates, R. 2024. Introduction to Responsible AI. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, 1114–1117. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703713.

Bainbridge, L. 1983. Ironies of automation. *Automatica*, 19(6): 775–779.

Baxter, G.; and Sommerville, I. 2011. Responsibility Modelling For Resilience. In *Proceedings of the Fourth Resilience Engineering Symposium*, 22–28. Paris, France.: Presses des Mines.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Bernstein, S. 2017. Causal Proportions and Moral Responsibility. In *Oxford Studies in Agency and Responsibility Volume 4*. Oxford University Press. ISBN 9780198805601.

Bishop, P.; and Bloomfield, R. 1998. A Methodology for Safety Case Development. In Redmill, F.; and Anderson, T., eds., *Industrial Perspectives of Safety-critical Systems*, 194–203. London: Springer London. ISBN 978-1-4471-1534-2.

Buhl, M. D.; Sett, G.; Koessler, L.; Schuett, J.; and Anderljung, M. 2024. Safety cases for frontier AI. arXiv:2410.21572.

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*, 279: 103201.

Cobbe, J.; Veale, M.; and Singh, J. 2023. Understanding Accountability in Algorithmic Supply Chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 1186–1197. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Coeckelbergh, M. 2016. Responsibility and the Moral Phenomenology of Using Self-Driving Cars. *Applied Artificial Intelligence*, 30(8): 748–757.

Commission, E. June 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Cooper, A. F.; Moss, E.; Laufer, B.; and Nissenbaum, H. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 864–876. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Da Silva, M. 2024. Responsibility Gaps. *Philosophy Compass*, 19(9-10): e70002.

Defence, U. 2025. Safety Management Plan 03. Safety Planning. https://www.asems.mod.uk/guidance/posms/smp03.

Dekker, S. 2018. *Just Culture: Restoring Trust and Accountability in Your Organization, Third Ed.* Abingdon, United Kingdom: Routledge. ISBN 978-1472475787.

Deo, Y.; Bonazzola, R.; Dou, H.; Xia, Y.; Wei, T.; Ravikumar, N.; Frangi, A. F.; and Lassila, T. 2023. Learned Local Attention Maps for Synthesising Vessel Segmentations from T2 MRI. In Wolterink, J. M. e. a., ed., *Simulation and Synthesis in Medical Imaging*, 32–41. Cham: Springer Nature Switzerland. ISBN 978-3-031-44689-4.

Edwards, L. 2022. Regulating AI in Europe: four problems and four solutions. https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf.

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. arXiv:2306.01917.

Food and Drug Administration (FDA) and Medicines and Healthcare products Regulatory Agency (MHRA). 2024. Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles.

Fulbrook, J. 2021. Reverberations from Uber v Aslam in personal injury claims? *Journal of Personal Injury Law*, 2: 59–67.

Grinfeld, G.; Lagnado, D.; Gerstenberg, T.; Woodward, J.; and Usher, M. 2020. "Causal Responsibility and Robust Causation". *Frontiers in Psychology*, 11.

Gunkel, D. 2012. *The machine question: Critical perspectives on AI, robots, and ethics*. Cambridge, MA 02142, United States: MIT Press. ISBN 978-0262017435.

Hart, H. L. A. 2008. 210POSTSCRIPT: RESPONSIBILITY AND RETRIBUTION. In *Punishment and Responsibility: Essays in the Philosophy of Law*. Kettering, Northamptonshire, U.K.: Oxford University Press. ISBN 9780199534777.

Hawkins, R. 2013. The Principles of Software Safety Assurance.

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). arXiv:2102.01564.

Hindriks, F. 2022. The problem of insignificant hands. *Philosophical Studies*, 179: 829–854.

Hitchcock, C.; and Knobe, J. 2009. Cause and Norm. *Journal of Philosophy*, 106(11): 587–612.

IEC. 2010. IEC-61508 Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems.

IEC. 2016. IEC-61882 Hazard and operability studies (HAZOP studies) – Application guide.

Independent. 2023. The backup driver in the 1st death by a fully autonomous car pleads guilty to endangerment. https://www.independent.co.uk/news/uber-aphoenix-maricopa-county-tesla-b2384012.html.

Institute, A. T. 2023. AI Safety Standards Hub. https://aistandardshub.org/ai-standards-search/. Accessed January 2025.

ISO. 2018. ISO-26262 Road Vehicles Functional Safety.

ISO. 2019. ISO-14971 Medical devices. Application of risk management to medical devices.

Joshi, I.; and Morley, J. 2019. Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care.

Kelly, T. 1998. Arguing Safety, a Systematic Approach to Managing Safety Cases. PhD Thesis, Department of Computer Science, University of York.

Kelly, T. 2004. A Systematic Approach to Safety Case Management. *SAE Transactions*, 113: 257–266.

Kim, T.; Perera, L.; and Sollid, M. 2022. Safety challenges related to autonomous ships in mixed navigational environments. *WMU J Marit Affairs*, 21: 141–159.

Koopman, P.; and Wagner, M. 2017. Autonomous Vehicle Safety: An Interdisciplinary Challenge. *IEEE Intelligent Transportation Systems Magazine*, 9(1): 90–96.

Kou, T. 2024. From Model Performance to Claim: How a Change of Focus in Machine Learning Replicability Can Help Bridge the Responsibility Gap. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1002–1013. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Lagnado, D. A.; Gerstenberg, T.; and Zultan, R. 2013. Causal Responsibility and Counterfactuals. *Cognitive Science*, 37(6): 1036–1073.

Lawton, T.; Morgan, P.; Porter, Z.; Cunningham, A.; Hughes, N.; Iacovides, I.; Jia, Y.; Sharma, V.; and Habli, I. 2024. Clinicians Risk Becoming "Liability Sinks" for Artificial Intelligence. *Future Healthcare Journal*, 11.

Leveson N., and Thomas J. 2018. STPA Handbook.

Lock, R.; Storer, T.; Sommerville, I.; and Baxter, G. 2009. Responsibility Modelling for Risk Analysis. In *Proceedings of the European Safety and Reliability Conference, ESREL 2009*. London: CRC press. ISBN 978-0-415-55509-8.

Luo, F.; Qian, H.; Wang, D.; Guo, X.; Sun, Y.; Lee, E. S.; Teong, H. H.; Lai, R. T. R.; and Miao, C. 2022. Missing Value Imputation for Diabetes Prediction. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Macrae, C. 2022. Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk. *Risk Analysis*, 42(9): 1999–2025.

Madaio, M.; Kapania, S.; Qadri, R.; Wang, D.; Zaldivar, A.; Denton, R.; and Wilcox, L. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1544–1558. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Matthias, A. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6: 175–183.

McDermid, J.; Nicholson, M.; Pumfrey, D.; and Fenelon, P. 1995. Experience with the application of HAZOP to computer-based systems. In *COMPASS '95 Proceedings of the Tenth Annual Conference on Computer Assurance Systems Integrity, Software Safety and Process Security'*, 37–48. Washington, DC 20036-4910 USA: IEEE.

Munch, L.; Mainz, J.; and Bjerring, J. 2023. The Value of Responsibility Gaps in Algorithmic Decision-Making. *Ethics and Information Technology*, 25(1): 1–11.

National Transportation Safety Board. 2019. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018, NTSB/HAR-19/03.

Nissenbaum, H. 1996. Accountability in a computerized society. *Science and engineering ethics*, 2(1): 25–42.

Nævestad, T.; Storesund, I. H.; and Phillips, R. 2018. How can we improve safety culture in transport organizations? A review of interventions, effects and influencing factors. *Transportation Research Part F: Traffic Psychology and Behaviour*, 54: 28–46.

OpenAI. 2025. Introducing ChatGPT.

Ozturk, B.; Lawton, T.; Smith, S.; and Habli, I. 2023. Predicting Progression of Type 2 Diabetes using Primary Care Data with the Help of Machine Learning. In *Medical Informatics Europe 2023*, 38–42.

P. Morgan. and et. al. 2023. *Tort Liability and Autonomous Systems Accidents*. Cheltenham, U.K.: Edward Elgar Publishing, Ltd. ISBN 978 1 80220 383 7.

Parliament, E. February 2023. Artificial intelligence liability directive briefing. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2023)739342.

Porter, Z.; Al-Qaddoumi, J.; Conmy, P. R.; Morgan, P.; McDermid, J.; and Habli, I. 2023. Unravelling Responsibility for AI. arXiv:2308.02608.

Porter, Z.; Zimmermann, A.; Morgan, P.; McDermid, J.; Lawton, T.; and Habli, I. 2022. Distinguishing two features of accountability for AI technologies. *Nature of Machine Intelligence*, 4: 734–736.

Rismani, S.; Shelby, R.; Smart, A.; Delos Santos, R.; Moon, A.; and Rostamzadeh, N. 2023. Beyond the ML Model: Applying Safety Engineering Frameworks to Text-to-Image Development. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 70–83. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

RTCA/EUROCAE. 2011. DO-178C Software Considerations in Airborne Systems and Equipment Certification.

Ryan, P.; Badyal, A.; Sze, S.; Hardin, B.; Bin Firoz, H.; Lewinska, P.; and Hodge, V. 2025. Safety Assurance Challenges for Autonomous Drones in Underground Mining Environments. In *Towards Autonomous Robotic Systems: 25th Annual Conference, TAROS 2024, London, UK, August 21–23, 2024, Proceedings, Part I*, 169–181. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-72058-1.

Ryan, P.; von Essen, M.; Shackley, L.; and McDermid, J. 2024. Bridging the Reality Gap: Assurable Simulations for an ML-Based Inspection Drone Flight Controller. In Ceccarelli, A. e. a., ed., *Computer Safety, Reliability, and Security. SAFECOMP 2024 Workshops*, volume 14989, 412–424. Cham: Springer Nature Switzerland. ISBN https://doi.org/10.1007/978-3-031-68738-9_33.

Ryan Conmy, P.; Mcdermid, J.; Habli, I.; and Porter, Z. 2023. Safety Engineering, Role Responsibility and Lessons from the Uber ATG Tempe Accident. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, TAS '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400707346.

Ryan Conmy, P.; Ozturk, B.; and Habli, I. 2023. The Impact of Training Data Shortfalls on Safety of AI-based Clinical Decision Support Systems. In Guiochet, J.; Tonetta, S.; and Bitsch, F., eds., *Computer Safety, Reliability, and Security*, volume 14181, 213–226. Cham: Springer Nature Switzerland. ISBN 978-3-031-40923-3.

Sartorio, C. 2007. Causation and Responsibility. *Philosophy Compass*, 2(5): 749–765.

Shepardson, D. 2023. Backup driver in 2018 Uber self-driving crash pleads guilty. https://www.reuters.com/business/autos-transportation/backup-driver-2018-uber-self-driving-crash-pleads-guilty-2023-07-28/.

Shoemaker, D. 2011. Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics*, 121(3): 602–632.

Smith, A. M. 2015. Responsibility as Answerability. *Inquiry*, 58(2): 99–126.

Sohal, K.; Mason, D.; Birkinshaw, J.; West, J.; McEachan, R. R. C.; Elshehaly, M.; Cooper, D.; Shore, R.; McCooe, M.; Lawton, T.; Mon-Williams, M.; Sheldon, T.; Bates, C.; Wood, M.; and Wright, J. 2022. Connected Bradford: a Whole System Data Linkage Accelerator. *Wellcome open research*, 7: 26.

Stahl, B.C. 2023. Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems. *Nature, Scientific Reports*.

Sujan, M. A. 2023. *Looking at the Safety of AI from a Systems Perspective: Two Healthcare Examples*, 79–90. Cham: Springer Nature Switzerland. ISBN 978-3-031-32633-2.

Sujan, M. A.; Furniss, D.; Grundy, K.; Grundy, H.; Nelson, D.; Elliott, M.; White, S.; Habli, I.; and Reynolds, N. 2019. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ health & care informatics*, 26(1).

Sujan, M. A.; Habli, I.; Kelly, T.; Pozzi, S.; and Johnson, C. 2016. Should healthcare providers do safety cases? Lessons from a cross-industry review of safety case practices. *Safety science*, 84: 181–189.

Tadros, V. 2018. Causal Contributions and Liability. *Ethics*, 128(2): 402–431.

Thompson, D. F. 1980. Moral Responsibility of Public Officials: The Problem of Many Hands. *American Political Science Review*, 74(4): 905–916.

Thompson, D. F. 2017. *Designing Responsibility: The Problem of Many Hands in Complex Organizations*, 32–56. Cambridge: Cambridge University Press.

Transformation Directorate. Jan 2025. Understanding AI regulation. https://transform.england.nhs.uk/ai-lab/explore-all-resources/understand-ai/understanding-ai-regulation/.

U.K. General Public Acts. 1974. Health and Safety at Work etc. Act 1974. https://www.legislation.gov.uk/ukpga/1974/37/contents.

U.K. Maritime and Coastguard Agency. 2022. Improving Safety and Organisational Performance Through A Just Culture . https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/286139/just_culture.pdf.

U.K. Ministry of Defence. 2017. Safety Management Requirements for Defence Systems 00-056 Part 1, Issue 7. Online: accessed 2022.

U.S. Food and Drug Administration. Oct 2021. Good Machine Learning Practice for Medical Device Development: Guiding Principles. https://www.fda.gov/media/153486/download.

Watch, H. R. 2023. How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers. https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net.

Weaver, B. W.; and DeLucia, P. R. 2022. A Systematic Review and Meta-Analysis of Takeover Performance During Conditionally Automated Driving. *Human Factors*, 64(7): 1227–1260. PMID: 33307821.

Wired. 2023. The Legal Saga of Uber's Fatal Self-Driving Car Crash Is Over. https://www.wired.com/story/ubers-fatal-self-driving-car-crash-saga-over-operator-avoids-prison/.

Yazdanpanah, V.; Gerding, E.; Stein, S.; Dastani, M.; Jonker, C.; Norman, T.; and Ramchurn, S. 2022. Reasoning About Responsibility in Autonomous Systems: Challenges and Opportunities. *AI and Society*, 38.

Zafar, M. 2024. "Normativity and AI moral agency". *AI and Ethics*, 4.

Zainuri, N.; Jemain, A.; and Muda, N. 2015. A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana*, 44(3): 449–456.