Artificial Intelligence 279 (2020) 103201

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective

Simon Burton [a], Ibrahim Habli [b,*], Tom Lawton [c], John McDermid [b], Phillip Morgan [d], Zoe Porter [e]

[a] Robert Bosch GmbH, Germany
[b] Department of Computer Science, University of York, United Kingdom of Great Britain and Northern Ireland
[c] Bradford Royal Infirmary and Bradford Institute for Health Research, United Kingdom of Great Britain and Northern Ireland
[d] York Law School, University of York, United Kingdom of Great Britain and Northern Ireland
[e] Department of Philosophy, University of York, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Article history:
Received 19 March 2019
Received in revised form 7 August 2019
Accepted 6 November 2019
Available online 9 November 2019

Keywords:
Safety
Autonomous systems
Artificial intelligence
Law
Ethics

ABSTRACT

This paper brings together a multi-disciplinary perspective from systems engineering, ethics, and law to articulate a common language in which to reason about the multi-faceted problem of assuring the safety of autonomous systems. The paper's focus is on the "gaps" that arise across the development process: the semantic gap, where normal conditions for a complete specification of intended functionality are not present; the responsibility gap, where normal conditions for holding human actors morally responsible for harm are not present; and the liability gap, where normal conditions for securing compensation to victims of harm are not present. By categorising these "gaps" we can expose with greater precision key sources of uncertainty and risk with autonomous systems. This can inform the development of more detailed models of safety assurance and contribute to more effective risk control.

1. Introduction

Autonomous systems disrupt established practices of system design, moral responsibility, legal liability, and safety assurance. By "autonomous system," we mean a system that makes decisions independently of human control. Autonomous decision-making capability can be understood as a spectrum. At the upper end of this spectrum, there are highly autonomous systems to which we delegate both the decisional process and the implementation of the subsequent action. At the lower end, there are advisory systems to which we delegate some or most of the decisional process, but the implementation of the recommended action is the responsibility of the human in charge. This paper considers the spectrum, but illustrates the ideas by considering both types of autonomous system in case studies.

The disruption that autonomous systems present can be explained in terms of 'gaps' that arise across the design and implementation process. We look at three gaps: the semantic gap; the responsibility gap; the liability gap.

The **semantic gap** arises because the normal conditions are not met for manufacturers, particularly designers, to provide a complete specification of the system. The semantic gap represents the difference between the implicit intentions

* Corresponding author.
E-mail addresses: Simon.Burton@de.bosch.com (S. Burton), ibrahim.habli@york.ac.uk (I. Habli), tom.lawton@bthft.nhs.uk (T. Lawton), john.mcdermid@york.ac.uk (J. McDermid), phillip.morgan@york.ac.uk (P. Morgan), zoe.porter@york.ac.uk (Z. Porter).

https://doi.org/10.1016/j.artint.2019.103201
0004-3702/

on the system's functionality and the explicit, concrete specification that is used to build the system. It is a risk in the design phase, but it is also an ongoing problem, with amendments to specification being required after implementation and deployment.

The **responsibility gap** arises because the normal conditions are not met for manufacturers, operators or users truly to deserve moral blame for the autonomous system's decisions, such as those that result in an accident or injury. The responsibility gap represents the difference between a human actor being involved in the causation of an outcome and having the sort of robust control that establishes moral accountability for the outcome.

The **liability gap** arises because the normal conditions are not met for manufacturers, operators, or users to be liable to pay compensation to those injured by an autonomous system. It is the risk that due to the complexity of the system, and its autonomous decision-making capability, that when it causes harm to others the losses caused by the harm will be sustained by the injured victims themselves and not by the manufacturers, operators or users of the system, as appropriate.

The gaps share the same three root causes: the complexity and unpredictability of the system's operational domain; the complexity and unpredictability of the system itself; and the increasing transfer of decision-making function from human actors to the system. These three issues also affect the safety assurance of autonomous systems. Not only is it much more complicated to provide a justification of the acceptable safety of autonomous systems, it is hard even to know what the required degree of confidence *is* within this new paradigm. The central, unifying argument of this paper is that measures to address each of the gaps above can inform the safety assurance of autonomous systems.

We ask the reader to note some clarifications. First, we place particular emphasis on autonomous systems that use machine-learning models rather than rule-based models. Second, for brevity, we refer to three kinds of human actor: the manufacturer, the operator, and the user. The term "manufacturer" should be understood as having a wide scope, including designers, developers, programmers, engineers, manufacturing companies. Third, this paper is primarily concerned with accident and injury as a result of an autonomous system's decision. Physical harm to humans is not the only risk incurred by autonomous systems, but it is an important one, and it is at the root of safety assurance, moral responsibility, and civil liability concerns.

The paper is structured as follows. In Section 2, we define and analyse the semantic gap, and its implications for safety assurance. In Section 3, we illustrate the semantic gap with two case studies: a highly automated driving system and a clinical advisory system. In Section 4, we consider the responsibility gap. We propose a methodology - the method of reflective equilibrium - for narrowing the responsibility gap during the design phase. In section 5, we consider the liability gap. We propose a solution - "tech neutrality" - for narrowing the liability gap; it also ensures that the law does not generate perverse incentives in technology adoption decisions. In Section 6, we build on this multi-disciplinary analysis to propose an approach to the safety assurance of autonomous systems.

2. The semantic gap

The Semantic gap is the gap between intended functionality and specified functionality – when implicit and ambiguous intentions on the system are more diverse than the system's explicit and concrete specification [9]. There are three sources of the Semantic gap. We refer to these as the three root causes throughout the paper.

The first root cause is the **complexity and unpredictability of the operational domain**. Autonomous systems typically operate within an environment that cannot be fully specified at design time. This is due to inherent environmental complexity at any given point in time, as well as continual, ongoing changes to the domain. Together these constitute an "open context," for which a complete specification is very difficult, if not impossible, to formalise.

The second root cause is the **complexity and unpredictability of the system itself**. Because of the computational techniques used, the system is inherently complex. It will also change continually through interactions within the domain. Systems that are expected to learn and to anticipate users' intentions may need to adapt to changes in these intentions over time, introducing risks not foreseeable during the design phase. In addition, the system has technical limitations, such as its incomplete "understanding" of the environment, which restricts its ability to react to certain situations. All of these factors arising from the complexity of the system make a complete specification problematic.

The third root cause is the **increasing transfer of decision function to the system**, whereby the human actor is either replaced completely (case study 1) or relieved of a substantial cognitive load (case study 2). This entails that new functions are introduced to the systems, including those that historically required human interpretation, ethical judgement and lawful behaviour. New functions are also introduced for manufacturers, operators, and users, such as the shift from active control of the system to passive, situationally-aware supervision. This further complicates the process of specifying intended functionality.

To some extent, the problem of ensuring that the system specification accurately represents "intent" is already a general problem in the assurance of complex systems. Various methods are typically used to confirm that the system specification meets user needs. These methods range from the systematic capture, analysis, and review of requirements to statistically representative field-based tests. But this is far more difficult against a backdrop of environmental complexity and unpredictability, along with increasing decision-making being transferred to the system. Here, safe system performance cannot always be guaranteed.

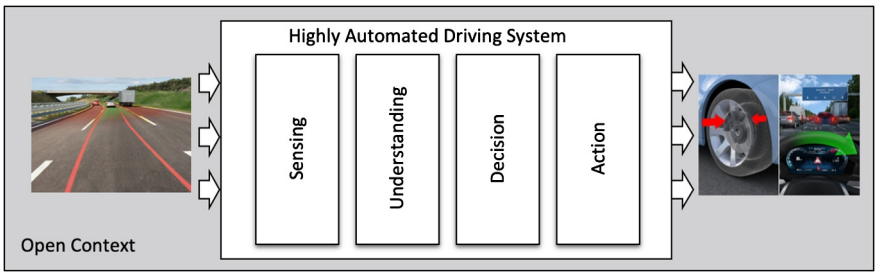

Fig. 1. Structure of a highly automated driving system.

3. Case studies

3.1. Case study 1 – automotive

3.1.1. Description of the system and current state of development

Within the last few years, many manufacturers have begun development of highly automated driving. Fig. 1 summarises the functional components of a highly automated driving system. ***Sensing*** components may consist of various direct sensor channels such as radar, lidar, and video cameras, but may also be extended to include indirect contextual information (e.g. from digital maps and vehicle-to-infrastructure systems). ***Understanding*** components interpret the current driving situation from the sensing inputs, processing raw sensory data to determine the current situation (i.e. vehicle position and trajectory, as well as the type, position and trajectories of other traffic participants). ***Decision*** components calculate driving strategies based on a set of driving goals (e.g. drive from A to B) and an interpretation of the current scene. ***Action*** components execute the driving strategy via the set of vehicle actuators (i.e. brakes, engine, steering column, etc.).

Recent advances in machine learning algorithms [25] and the availability of increased computing power mean that the systems themselves are more able to solve the "Understanding" and "Decision" tasks in unrestricted operational environments. Deep neural networks can make sense of unstructured data using efficient computations in real-time [21]. By providing enough training data, the models learn to identify and classify objects such as vehicles and pedestrians with accuracy rates that can surpass human abilities [59].

At present, two strands of development exist for highly automated driving. The first is SAE Level 3: Conditional Driving Automation Systems. These take over control of the vehicle whilst driving on highways [50]. Here, the human driver must be available to resume control in the case of a system failure or when the boundary of the operational design domain has been reached. These systems are an evolution of current driver assistance systems. There is therefore a degree of familiarity with their capabilities and their technical limitations.

The second application of automated driving is SAE Level 4: "High Driving Automation" systems for urban automated driving. Here, the system takes over Dynamic Driving Tasks (DDT) and Object and Event Detection and Response (OEDR) in highly complex environments, and has fall-back functions in case of system malfunction. This is a more revolutionary approach to mobility. It may take the form of not just classical passenger cars but also new classes of vehicles, such as driverless shuttles or delivery vehicles. These vehicles will travel at significantly lower speeds than with SAE Level 3, but their technical and safety assurance challenges are far greater.

3.1.2. How the system illustrates the semantic gap

The transition from hands-on (SAE Levels 1-2 [50]) of driver assistance to hands-off (SAE Levels 3-5) highly automated driving illustrates the Semantic gap. SAE Level 4 in particular shows how the three root causes described in Section 2 contribute to the Semantic gap.

First, urban environments are "crowded", complex, unpredictable operational domains with a high number of traffic participants of diverse capabilities, shapes and sizes, speeds and trajectories. It is also anticipated that this environment will change over time, as new modes of inner-city transport are introduced, and as human behaviour adapts to the presence of highly automated vehicles. It is therefore impossible to create a complete specification of the operating environment for urban automated driving.

Second, the systems themselves need to be more complex to manage the driving task in urban environments. Heterogeneous sensing channels are required to counteract the individual weaknesses of each sensor type (e.g. camera, radar etc.). The inputs of these sensors must be consolidated and interpreted to provide a model of the environment that can be used to implement an optimal driving strategy. The algorithms too must become more complex to enable the system to interpret the environment, to predict the intentions of traffic participants, and to make optimal decisions as to which action is required to minimise the overall risk to the vehicle occupants, other road users and the environment.

Machine learning techniques such as Deep Learning (enabling the system to "make sense" out of the unstructured data that results from the complex and unpredictable environment) and Reinforcement Learning (enabling the system continually to optimise a function based on stimuli collected in the field) might appear to present a convenient technical solution to the semantic gap. They seem particularly well-suited to learning functionality that cannot be easily specified using traditional

procedural means (if X happens, then do Y). But there is a catch. Machine learning functions do not deliver clear-cut answers. For example, for a given video frame, they might classify the probability of a pedestrian inhabiting a certain portion of the picture as 83%, but in the very next frame – which for humans is imperceptibly different to the last – they may "misclassify" the same object as only 26% probability of being a human and 67% probability of being a road sign. In addition, the processes which lead to these decisions are difficult to decipher. These attributes result in a paradox or "no free lunch" effect, where the problem of deriving a suitable specification of the intended behaviour is instead transferred to the problem of demonstrating that the implemented (learned) behaviour meets the intent.

The third root cause that contributes to the semantic gap with SAE Level 4 systems is the significantly increased transfer of decision function. Because of the lack of a "backup driver" to take over in critical situations, the concrete specification of the system must inform what the system itself should do under all situations, even when a completely safe state may not be able to be reached. This is a huge task, involving considerable uncertainty. Due to the open context and the complexity of the decision algorithms it may not be possible to even predict which decisions the vehicle would take under certain circumstances.

3.1.3. Implications for safety assurance

The dominating challenge facing the safety assurance of highly automated driving systems is the derivation and validation of adequate system safety requirements and the demonstration that these will be fulfilled in all feasible situations, including those that have traditionally been handled by a human driver. This results in a different class of safety requirements to those previously considered in the industry [50]. For example, a higher level of component reliability is required, because the system cannot be simply deactivated by a human driver upon detection of a component hardware fault. At a functional level, an approach to demonstrating the correct interpretation of the current driving situation, previously made by a human driver, is required so that dangerous driving situations are avoided as far as possible.

The conditions for acceptable functional safety for passenger vehicles are set by the international functional safety standard for road vehicles ISO 26262 [27]. This standard is limited to hazards, i.e. sources of harm, caused by the vehicles' malfunctioning behaviour. This remains necessary to ensure that the hardware is reliable and the implementation of systems is fault tolerant. But extensions to the standard to accommodate autonomous vehicles, in particular, the "Safety of the Intended Functionality" (SOTIF) approach, are currently focused on driver assistance, not highly automated driving systems [28]. As a result, additional approaches must be developed.

The ISO 26262 standard requires the development of a safety case: a valid, evidence-based justification for a set of claims about the safety of a system for a given function over its operational context [26]. In the context of the systems in this case study, this safety case should provide a structured argument, based on "first principles," that the driving function is safe for all conditions that meet the assumptions on the target domain. It must justify the acceptable level of residual risk[1] associated with this function. This justification will be partly based on the capability of the system architecture itself to minimise the risk given the complexity and unpredictability of the domain, sensing errors, and component insufficiencies [14]. The first fatal accidents caused by vehicles operating in this driving mode have highlighted the need to discuss and reach consensus on acceptable residual risk for such systems [43] [44]. This wider discussion should also include the potential of the systems to significantly increase road safety [2][18].

3.1.4. Suggestions for reducing the semantic gap

A number of approaches are currently being considered to close the semantic gap during the design of highly automated cars. These approaches target the gap's three root causes.

The complexity of the operating environment is reduced by limiting functionality to well-defined scenarios for which a clear understanding of the safety risks and system capabilities already exist. In the case of SAE Level 3 systems, this involves certain stretches of highway under limited weather conditions and functional constraints such as no overtaking. For SAE Level 4 systems, operation is restricted to geo-fenced areas of urban environments for which highly detailed maps and validation data exists [62]. These restrictions come at the expense of a reduction of the intended function.

The complexity of the systems is managed by limiting the deployment of machine learning algorithms to well-defined and constrained functionality with restricted safety impact. This includes using parallel approaches to sensing and plausibility checks. It is also planned to make increased use of infrastructure, such as the transmission of traffic signal status, to allow for more robust approaches to environmental sensing. In this way, the burden of validating the machine learning components is reduced. But reducing the integrity requirements on machine learning components through functional redundancy comes at the price of many more system components and overall cost.

In current systems, the delegation of decision function to the systems is also reduced. This can take a number of forms, including the requirement for driver supervision in Level 3 highway automation, the use of a safety driver when testing Level 4 urban automated driving, or dropping out into a safe state such as stopping at the side of the road in the case of ambiguous and potentially critical situations. These measures result in restrictions in the intended function.

It should be noted that addressing the semantic gap will not happen solely in the design phase. It will be an ongoing process, with amendments to specified functionality being made after it is understood how the systems operate "in the

[1] The risk that remains once all risk reduction measures have been taken.

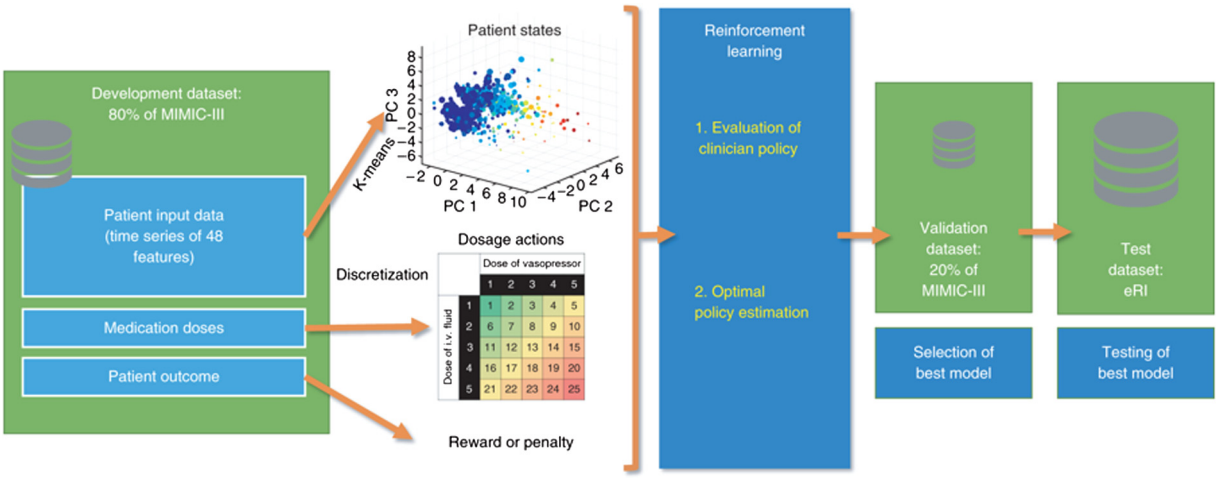

Fig. 2. Data flow of the AI Clinician [30].

wild," and once differences between usage in different cultures has been more fully understood, including how attitudes to acceptable safety vary between countries, etc. The compromises described above may be iteratively relaxed as more validation evidence is generated, technological advancements are made, and a stronger and a more dynamic overall safety case is achieved. This overall argument for safety must address the avoidance of hazards caused by the automated driving function itself, reactions to the failures of other systems in the vehicle, the prevention or mitigation of misuse of the driving system, and the reaction to critical driving situations caused by other road users and agents.

As yet, no consensus exists for how to reduce the semantic gap and structure a compelling safety case for highly automated driving. Governments and regulators are currently in the process of consultation to define the relevant legal frameworks [32]. Much clarification is still needed. There is therefore a strong need for a common language for expressing technically feasible ethical and legal expectations on such systems. Methods to achieve this are proposed in Sections 4 and 5.

3.2. Case study 2 – clinical

3.2.1. Description of the system and current state of development

The last few years have also seen the development of autonomous systems in the clinical domain. One example is "AI Clinician," which is an advisory system in the treatment of sepsis, based largely on a Reinforcement Learning agent (Fig. 2). It is designed to operate in the well-defined area of Intensive Care Unit patients fulfilling the Sepsis-3 criteria [30]. These are also the most unwell patients, with a 90-day mortality in the region of 20%.[2] It provides guidance on treatment. Whether or not to implement the system's recommendation is up to the human clinician. The AI Clinician can only make its decisions based on data entered into the Electronic Patient Record (EPR). The system was trained using the MIMIC-III dataset and uses a Markov Decision Process (MDP), which is 'memoryless' and ignores the individual patient's history.

Sepsis is associated with at least 1 in 20 deaths in England [38] and is a leading cause of mortality worldwide [17] [53]. Best practice for treatment is currently rapid administration of antibiotics to treat the infection, intravenous fluids to correct hypotension (low blood pressure), and vasopressor medications if the patient's blood pressure remains low [33] [22]. As both fluids and vasopressor medications can be used to improve blood pressure, there remains debate over the importance of each [52]. "Traditional" strategies favour fluids but there is some evidence that large volume fluid use can increase mortality. But vasopressors are powerful medications that, in the UK at least, require the patient to occupy a high-resource ICU (Intensive Care Unit) or HDU (High Dependency Unit) bed. As patients are on ICU, fluids and vasopressors are both easily available.

The AI Clinician aims to guide the use of fluids and vasopressors in septic patients and does not provide guidance on any other aspect of care; this narrow focus is probably of benefit.

The AI Clinician uses 48 variables to assign the patient to one of 750 states, and operates over a 4-hour window. It is therefore able to be much more responsive to more subtle changes in the patient's status than is possible for a human clinician and bedside nurse. Its recommendations are based on a 5-point scale for each of fluid and vasopressor prescription, with one zero point and 4 quartiles of dosing. The exact point chosen within a quartile would presumably be

[2] The definition of sepsis is hard to pin down, and the Sepsis-3 definition represents a narrower definition than previous ones which threatened to label almost all hospital patients as "septic". In hospital, the word is generally used in a more narrow sense than in the literature, but patients on ICU fulfilling the Sepsis-3 criteria will almost always be considered "septic" by doctors.

up to the bedside nurse with advice from human clinicians. In general, it appears that the advisory system-generated policy recommends more vasopressor use, and less fluid use, than the patients actually received.

3.2.2. How the system illustrates the semantic gap

As with the previous case study, the three root causes of the semantic gap described in Section 2 relate to this clinical advisory system.

First, the operational clinical domain for this system is highly complex, including sepsis itself, the presence of factors such as new treatments, new diagnoses, new bacteria and viruses, as well as differentials in patient care at earlier time points. There is also considerable debate within the medical community as to what constitutes best practice in the treatment of sepsis, variation of practice between individual clinicians, and a gap between clinician advice and what actually happens to the patient, given that the bedside nurse has some authority and is in direct control of the fluids and vasopressors. This complexity of the domain makes it very difficult to formalise intended functionality into specific requirements.

Second, there is the complexity of the AI Clinician system itself. As with the system in the first case study, there are the technical limitations of the system (e.g. reward hacking in reinforcement learning [1]). The AI Clinician can only make its decisions based on data entered into the Electronic Patient Record (EPR). By contrast, human clinicians can respond to a variety of inputs when making their decisions. The system also uses a MDP, which only looks at the present state of the patient, and thus is quite unlike human decision-making. In addition, there is a great deal of inertia the clinical decision-making process. It is highly likely that the patient's fluid/vasopressor state will vary little across time, and this may be a positive thing. An autonomous system would not have the same inertia unless learned, which is not possible with the methods used, or specifically programmed in. The implicit, intended functionality remains broader than what is possible to specify explicitly as functionality.

Third, there is the transfer of the decision function. Though this is an advisory system and the final decision is made by the human in charge, busy clinicians may come to over-rely on the system and not question its recommendations. A human must develop their own opinion on the correct course of action, rendering the system less useful, or blindly accept its recommendations, making it no longer advisory. It is also still difficult to ensure the process leading up to the system's recommendation mirrors what an ideal human would do. This is hard for two reasons. Because human clinicians are not MDP decision-makers, and because a clinician or nurse will be able to explain their general decision making process, there is a risk of a lot of this being "post hoc" rationalisation rather than a genuine opening of the human black box. While the developers of AI Clinician have been able to show that the factors important to the system were similar to factors expert clinicians would consider important, the transfer of decision function does affect the feasibility of providing a complete specification of intended functionality.

3.2.3. Implications for safety

The ideal method of testing AI Clinician would be "On Policy Evaluation." This involves actually following the system's recommendations and monitoring patient outcomes accordingly. These outcomes should be compared against a control arm of standard treatment. But this method is logistically and ethically challenging, and may not even be possible given that doctors and nurses may lose equipoise when the system recommends an action they consider unwise.

Therefore, testing of the AI Clinician to date has relied on "Off Policy Evaluation," where an attempt is made to judge the *expected* value of the system's recommendations against the actual decisions made and acted on by the clinical team. In order to do this, the study authors have tried to formalise the human clinician decision-making process as a MDP. This means treating each clinician decision as a choice based solely on the present state of the patient (i.e. ignoring patient history and any non-EPR data), with the assumption that any such patient is equivalent, and the actions taken by clinicians would be spread randomly (weighted) across the real actions taken by clinicians for each such patient. This formalisation of the human clinician decision-making allows it to be compared against the system. But ultimately, clinical decision-making is more complex than this and the formalisation carries the risk of mis-characterising the human decision-making process. Indeed, it is in trying to encode and specify the intention (treat and cure patient) as the taking of actions that have a predicted high probability of moving towards higher-scoring positions in state space, that the semantic gap emerges. The intended functionality requires something more diverse than this specification.

Fundamentally this remains a difficult problem for the safety assurance of AI Clinician – how to test the system's advice without following it, given that following it might be dangerous, resulting in serious physical harm to patients.

3.2.4. Suggestions for reducing the semantic gap

As with the automotive case study, approaches that narrow the semantic gap with this clinical advisory system target the gap's three root causes.

The complexity of the operational domain is reduced by ignoring aspects of the decisional process, such as limiting the number of informational inputs compared to those received by human clinicians. But this may result in unintended consequences (e.g. 'insensible losses' of fluid cannot be electronically recorded, which may lead to a system suggesting more fluids when a clinician could tell that the patient is already 'waterlogged'). Clinical decision making integrates history, physical examination, and recorded numbers, but only the last of these is available to the system.

The complexity of the system is managed by control over the formalisation process, in particular, as above, the representation of clinician decision-making as an MDP. One of the advantages of this is that the system can more easily be

explained. But, as we have seen this is not the best picture of how clinicians actually make decisions about the treatment of sepsis.

The delegation of decision function is also restricted to the degree that the human in charge is the final, authoritative decision-maker. But, as we have seen, this intended restriction may be ignored in practice because of human over-reliance on or misuse of the advisory systems.

Thus, as with the first case study, reducing the semantic gap is currently achieved by restricting the intended functionality. But this may inhibit the benefits that the systems can bring to reducing mortality due to sepsis. Sections 4 and 5 of this paper look at other measures to address this gap.

4. The responsibility gap

4.1. Identifying the responsibility gap

Moral responsibility is integral to practical ethics. It concerns what makes a person deserve moral praise or blame for some outcome. A person can be morally responsible in the absence of legal responsibility and legal obligations, and vice versa. This section concerns only moral responsibility. Determining who we can justifiably hold morally responsible for harm and injury due to system behaviour will likely be a precondition of public trust in autonomous systems.

The three root causes of the semantic gap also underlie the responsibility gap. Because of the complexity and unpredictability of the operational domain, there is increased risk that the system could cause injury; it is also more likely that the system will encounter what would be an 'ethical dilemma' for a human carrying out the same function. Because of the transfer of decision function to the system, its behaviour is no longer under the direct control of manufacturers, operators, or users. And due to the system's complexity, its decisional process is largely unexplainable, even by experts. It is generally agreed that people do not truly deserve blame for actions they have no control over, or about which they are ignorant. As such, though manufacturers, operators, and users are causally involved in the process, this is not sufficient for robust, retrospective moral accountability for many of an autonomous system's actions. The responsibility gap occurs across all phases; this paper analyses the specific responsibility gap that arises from behaviour due to the system's design (i.e. from the semantic gap). We propose a procedure - the reflective equilibrium - inspired by political and moral theory – that can be used in the design phase to address this specific responsibility gap.

With traditional complex engineering systems, responsibility gaps may arise, but most of the time, the conditions are in place to hold either the manufacturers, the operators or the users morally responsible for accidents and injuries consequent on their design, engineering, or use. This is because there is no handover of the decision-making function, and system behaviour more directly represents human intentions; moreover, internal processes causing the behaviour are, for the most part, intelligible models. But with autonomous systems, moral responsibility for harm-causing behaviour may not be ascribable even in principle to the manufacturer, the operator or the user.

The responsibility gap was first characterised by Andreas Matthias as follows: "*... there is an increasing class of machine actions, where the traditional ways of responsibility ascriptions are not compatible with our sense of justice and the moral framework of society because no one has enough control over the machine's actions to be able to assume responsibility for them.*" [35].

There is a growing body of literature on responsibility gaps [35][23][10][39][46][24]. Because of the severity of risk and the importance of accountability in the military domain, the notion is often discussed with respect to autonomous weapons and meaningful human control [55] [51][42]. But responsibility gaps are not exclusive to autonomous weapons.

Our treatment highlights two dimensions to the responsibility gap: the Control condition and the Epistemic condition. This is how the gap is starting to be characterised in the philosophical literature [24]. The Control and Epistemic conditions are sometimes called the 'Aristotelian conditions': they can be traced back to Aristotle's treatment of moral responsibility [3][19][10]. These can be considered as necessary conditions: moral responsibility only obtains if the two conditions are met:

- **Control condition**: the person must have relevant control over the action, such that the action adequately represents or reflects the person's intentions or desires.
- **Epistemic condition**: the person must have had relevant knowledge and understanding of the action, and its likely consequences.

There is substantial philosophical debate about the precise content of these two conditions, but the present discussion can take place at a higher level of generality. To note, the Epistemic condition does not excuse negligent ignorance.

4.2. Challenges to the conditions: how the case studies illustrate the responsibility gap

In the **first** case study, we have a highly automated driving system. Here, the Control condition is weakened in general because it may not be possible for a human to take over control of the system, and because the system's actions may not reflect intentions on it. The Epistemic condition is weakened in general because the system's inherent opacity due to its machine learning algorithms mean that its decisions are often inscrutable to manufacturers *post hoc*.

In this paper, we focus on the specific threat to moral responsibility arising from the process of design specification. As such, the responsible party considered here is the manufacturer, where this term has the wide scope highlighted in Section 1. The Control condition would require that behaviour resulting from the driving system's specified functionality represents the manufacturer's intentions. But it is difficult to achieve this when it is not possible to provide an adequate, explicit formalisation of those intentions. The Epistemic condition would require that the manufacturer has relevant understanding of driving behaviour that should result from the system's specified functionality. But not only is it hard to foresee how an inherently complex, decision-making system will behave in the unpredictable operating environment, there is also ethical and legal uncertainty about what the system should do in boundary cases, where typically the human decision-maker would exercise some individual moral judgement.

Some of the ethical concerns fielded in respect of highly autonomous vehicles are an indirect result of the difficulties raised by the semantic gap [6]. The Trolley Problem is often cited in this respect - somewhat misleadingly in the authors' opinion. The Trolley Problem puts forward the following dilemma: a faulty trolley is headed towards five people who will be killed by the collision, but the driver could steer the trolley onto another track where there is only one person, a workman [20][56]. In recent years, this thought experiment has been applied to autonomous cars: if the runaway trolley were in fact a highly automated driving system, what should it do? [6][19][34][39].

The Trolley Problem is an unhelpful analogy from a design perspective. In practice, it is not something that would be specified as part of the intended function at all. It would be avoided through technical requirements such as predictive and conservative driving. The vehicles are being designed to avoid accidents when they can and, where this is not possible, to adopt other strategies, such as the minimisation of kinetic energy on impact. Indeed, their particular sensing abilities mean they would be able to use the braking system in better time than human drivers to minimise the risk of accident [36]. This conclusion is reflected in the German Ministry of Transport and Infrastructure Ethics Commission rules [16] on dilemma situations: they should be avoided by design, and the public should be made aware of behavioural traits and performance limits of highly automated driving vehicles on the road.

The Trolley Problem is also problematic from a philosophical perspective. The thought experiment is intended to apply to an ideal, all-things-equal scenario, abstracted from all other contextual factors. But the design challenge is precisely that highly automated driving systems are not being designed for all-things-equal operational environments [45]. They are being designed for a dynamically changing environment, in which small situational changes are relevant to critical decisions. Thus the design problem and the Trolley Problem are disanalogous. And because of the complexity and unpredictability of the operational environment, it is not epistemically possible for manufacturers to imagine in advance all the combinations of factors and fine distinctions that might affect the permissibility of the action.

But while the Trolley Problem does not express a realistic uncertainty for manufacturers, there are other examples in which ethical judgements would be made by human drivers, and it is unclear how this should be formally specified in the system. An often discussed dilemma is maintaining the flow of traffic even when this is in violation of a traffic law, such as crossing a solid line in the middle of the road to avoid an obstacle in the lane. In some cases, keeping to a fixed set of rules would not be the anticipated behaviour of other drivers; it could lead to a risk of a rear-end collision.

In the **second** case study, we have a clinical advisory system. Because the final decision here is implemented by a human, there is only a partial transfer of decision function, and more human control, though as we have seen humans may over-rely on the system over time. The responsibility gap that is incurred by such systems is narrower, but it does still arise. There are general epistemic problems in explaining the decisional process of the inherently complex system. And there are more particular challenges to the two conditions, some of which arise specifically during the design phase. As above, in terms of the Control condition, it is difficult to make sure the system's recommendations represent intended functionality when it is not possible fully to formalise intention. In terms of the Epistemic condition, it is hard to know or understand what the intended functionality should be. In the case of AI Clinician, this is because what constitutes best practice depends on the interactions between clinicians and nurses, and the fact that best practice changes, both in terms of an evolving understanding of optimal behaviour (e.g., a gradual move away from fluids towards vasopressors) and in terms of environmental changes (e.g., new diseases and treatments) that serve to change what is "optimal".

4.3. The method of reflective equilibrium: a suggestion for reducing the responsibility gap

A partial solution to the problem of the responsibility gap incurred during system design can be drawn from the method of the "reflective equilibrium," as articulated primarily by the philosopher John Rawls [13][49][48]. While others have posited this method as an approach to designing the internal reasoning capacity of ethical advisory systems [63], our contribution is to propose an adaptation of this method to navigate some of the problems arising from the semantic gap - in particular, so that manufacturers have more responsibility-grounding control with respect to harm due to the system's design.

The method of reflective equilibrium is particularly helpful to decide the best course of action where there is uncertainty [16]. It is an inductive reasoning process that involves agents making specific, context-sensitive judgements (e.g. judgements about the system's desired functionality), on the one hand, and evaluating these against a wider set of moral principles and non-moral beliefs, on the other, and revising each until there is an acceptable coherence among them [13]. We suggest that the process can offer greater clarity on intended functionality, including where 'ethical dilemmas' are concerned. This can enable a more complete system specification.

Here is a sketch of how it might work in the **first** case study. Competent moral judges, with sufficient technical understanding as well as experience of urban driving environments, and who would not gain personally from any particular specification agreed, could work in multi-disciplinary design teams and make considered judgements about ethical system behaviour in actual and hypothetical "boundary" scenarios (e.g., exceeding the speed limit to avoid an obstacle). Boundary scenarios are important because it is typically on the boundaries of expectations that hazards arise. Those judgements could be considered against their own and others' different moral and non-moral beliefs (e.g. beliefs about the practical consequences of excessively cautious driving and beliefs about what is technically or computationally feasible). The process involves achieving equilibrium between the judgements of various people involved in the development of the systems; legal experts could also be included here.

The hypothetical boundary scenarios that warrant special attention can be identified during simulation of the system in the design phase. Integral to this approach is the automatic detection or generation of such scenarios based on a definition of risk, such as a high likelihood of harm no matter what the system's action. During simulation, these scenarios can be replayed with various parameters, to explore the expected boundaries of the system behaviour. Indeed, the process can lead to a set of reference situations that better determine the expected range of system behaviour in critical scenarios. Because judgements are evaluated against technical feasibility at the start of the process, far-fetched dilemmas such as the Trolley Problem would be ruled out early on, but more realistic dilemmas would be incorporated. Where certain judgements are deemed unacceptable, because they are incompatible with other deeply held beliefs and principles, revisions to the initial judgements can be made. At times, revisions to the principles may be made instead. These revisions may include, for example, limiting the intended functionality or accepting a higher risk or uncertainty.

Applied to the **second** case study, the method of reflective equilibrium might work as follows. Competent moral judges, again with sufficient technical understanding as well as clinical knowledge of sepsis and ICU settings, can work in multi-disciplinary teams to identify the most salient open questions (e.g. patient history) and features of the domain (e.g. the different aetiologies of sepsis). They can evaluate these judgements against different moral and non-moral beliefs (e.g. beliefs about best practice and beliefs about technical feasibility). The method can also be used to gauge foreseeable misuse - such as busy clinicians over-relying on the system, or being dis-incentivised to override the system's recommendations. Moreover, the decision procedure may help to establish principled positions on when such uses would be unacceptable given the clinician's retained responsibility for patient safety. Overall, reflective equilibrium offers a methodology by which to navigate the complexities of the domain and the expectations on the system.

Because the method of reflective equilibrium enables manufacturers to have more rational control and understanding during the design phase, thereby addressing the Control and Epistemic dimensions of the responsibility gap in this phase, it helps to secure more robust moral accountability for behaviour arising from the system's specified functionality. It also facilitates consensus on 'acceptable safety' for different autonomous systems, which can ground a stronger justification for the specified functionality of the systems. But some caution is also required. First, the reflective equilibrium may help to narrow the responsibility gap arising from the semantic gap, but as it has been discussed here, it does not address the further problem of moral responsibility for outcomes that are traceable to events and interventions post-design [9]. The second note of caution is pragmatic: the procedure would need to be robustly managed, with workable limits, to ensure that real progress is made in the timeframes appropriate for the rate of change of systems. Third, it would ideally be overseen by regulatory bodies for autonomous systems - and this is likely to be a significant change for the majority of regulators.

5. The liability gap

5.1. Identifying the liability gap

There is a liability gap for harms caused by autonomous systems. This gap partly emerges from the replacement of human actors in general; it also results from this combined with the inherent complexity of the system and the complexity of the operational domain. This combination may lead to the loss falling on the victim of the harm. As well as describing the liability gap, and illustrating it with the two case studies, this section attempts to solve this gap, and introduces the concept of "tech neutrality" in the law of tort.

5.1.1. Tort law

There are a number of legal systems in the World, here we examine the English common law as a case study. This legal family is found throughout the English speaking world, for instance in the United States, Singapore, and Australia. Within this legal family [64], liability law has the same structure and language [60]. The problems, and liability gap, encountered in England and Wales in relation to liability for harms caused by autonomous systems are typical of all common law jurisdictions.

Harm resulting to another from the use of a product or system is typically resolved through the law of tort. Tort law is concerned with civil liability for wrongs, and typically redresses the harm through financial compensation, (called damages), or less often injunctions to prevent continuing harm. We need to distinguish tort from crime. Most harms inflicted by autonomous systems will not be crimes if they had instead been committed by a person. For instance not all who cause the deaths of others are criminalised, even if they are at fault. The standard of fault required for criminalisation is much

higher than the standard of fault required to be civilly liable to pay compensatory damages to the victim. Crime is also rarely concerned with unintentional wrongs that result in non-fatal harm. Thus this paper only considers civil law.

Tort law, whilst it also has a regulatory function, is not the same as government regulation. Claims are brought against the wrongdoer by the parties who are injured, and not by the state. Even if a regulatory body has approved a product, it does not mean that there is no liability in tort if harm results from its use [58]. Tort looks at redressing harms, whilst also having regulatory force [31].

Sometimes the development of novel products can outpace the capacity of legislative and administrative bodies to produce new regulatory codes for such products, or alternatively they may decide such codes are inappropriate, and leave regulation solely to the ordinary law of tort [41]. Ordinarily tort, not regulation, provides redress for harm. Thus to oust tort is to oust compensation for harm. The broadest, and most commonly applied tort is the tort of negligence.

5.1.2. Harms caused by persons

Before the liability gap in tort generated by the adoption of autonomous systems in place of human workers becomes apparent, we need to briefly deal with the situation of where a person causes actionable harm to another. Not all harms are actionable, but here we are concerned with personal injury, which is.

If A negligently injures B, A has committed the tort of negligence and is liable to compensate B for his loss. C may also be liable to compensate B for A's tort via a form of strict liability called "vicarious liability", provided firstly, there is a relationship between A and C sufficient to trigger the doctrine; and secondly, A's tort must be sufficiently connected with that relationship to render C vicariously liable for the tort [40]. Vicarious liability typically occurs in the context of employment, making an employer vicariously liable for the torts of its employees, provided the employee's tort was closely connected with their employment. However, it also applies to a range of other relationships, particularly where C has control over A - that is the ability to tell A what to do, and perhaps also how to do it, and A's work is integrated into C's organisation [41]. A broad approach is taken to close connection [5]. It goes without saying that a crash with another vehicle caused by the poor driving of an employee required to drive as part of their employment would result in vicarious liability for the employee's negligent driving, as would the death of a patient resulting from the poor treatment of sepsis by a medical practitioner. In such cases both A and C are liable to the victim B (or their estate) for A's negligence.

Where a person, A, commits a tort against B; C may also be liable to B for C's own negligence where C's own negligence results in A committing a tort against B. This is separate from vicarious liability, and it for instance applies where C owed B a duty of care to properly select, train, and monitor A, and where C fails to meet the standard expected in doing so, and this failure resulted in A committing the tort against B [5]. B's claim against C is called a direct duty claim. This direct duty claim is based on C's own negligence. However, the vicarious liability claim against C is easier to sustain in that unlike the direct duty claim you do not need to prove that C owed B a duty of care, nor do you need to prove C's fault. Now that we have an overview of the liability systems in place for individuals, the liability gap created by the use of autonomous systems may now be identified.

5.1.3. The liability gap

The liability gap results from the replacement of employees with autonomous systems. This gap occurs since whilst the system can harm another in ways similar to an employee, unlike an employee it cannot commit a tort. This is since systems or products are not legal persons, meaning that they are not subject to rights or duties in law.

That autonomous systems are not persons results in a liability gap, when compared to the scenario of A, B, and C above. Where the user (C) uses an autonomous system (Φ) in place of A, and Φ causes harm to B, B of course cannot sue Φ (since Φ is not a legal person), however, it also means that B cannot sue C vicariously, since vicarious liability requires a tort, and since Φ is not a legal person it cannot commit a tort. This is the case even if the action of Φ would be a tort if it were committed by a person.

It may be possible to construct, by analogy to the employment situation, a duty of care owed by C to B to take reasonable care to select a suitable system (Φ), and to take reasonable care to maintain, update, and monitor its operation. However, given the need for C's fault for this claim to succeed this direct duty claim will not provide for liability in all cases where vicarious liability would apply if Φ were replaced with a human (A). Indeed given the complexity of autonomous systems, and the often black box situation associated with its AI components, the user of Φ is rarely likely to be at fault provided they follow the manufacturer's guidance.

The existence of the semantic gap will also mean that unexpected harm may result from the use of the system. Further, the employing enterprise will only be directly (as opposed to vicariously) liable where the type of harm that results from the enterprise's own negligence is foreseeable.

For the user C, using Φ in place of A reduces their exposure to liability, at B's (the victim's) cost. However, the liability gap of the enterprise that uses the technology in place of employees, may be partly offset by the fact that there can additionally be liability on the part of the manufacturers of Φ. Although, as we will see below there are significant limits to this liability, and it does not fully account for the liability gap created by the absence of vicarious liability.

5.1.4. Claims against manufacturers do not close this gap

Unlike where A is human, with Φ, there are potential claims against Φ's manufacturer. The first claim is under the UK's Consumer Protection Act 1987, which implements an EU directive [11]. The Act introduces liability without fault in certain

circumstances, which is caused by a defective product. Section 2(1) states: "... where any damage is caused wholly or partly by a defect in a product, every person to whom subsection (2) below applies [i.e. the manufacturer and others] shall be liable for the damage."

It's important to note that there is no need for fault here. The claimant relying on the Act does not need to prove that the defendant was negligent. The Act refers to products [11] and this is unlikely to include software unless it is supplied in a physical medium [32]. This is problematic in claims involving machine learning if their software is subsequently updated over the internet, and for cloud based products.

Even though the claimant does not need to prove fault, they need to prove that the product is defective. The product is deemed defective "if the safety of the product is not such as persons generally are entitled to expect..." [32]. Where a product meets the safety standards imposed by the relevant regulatory regimes, it is difficult to argue that the product is defective [32].

Where a product is not only inherently dangerous but known to everyone to be so, (such as hot coffee), its danger has to be factored into the public's reasonable expectation for the purpose of defectiveness. This will also be the case with autonomous systems.

Key for claims against a manufacturer when dealing with the semantic gap, and machine learning, is a defence [32] the defendant is not liable if the defect comes into existence after he parts with the product. Thus if the defective behaviour comes into existence after the manufacturer parts with the product (unless it is the learning design itself that is defective) then it is likely that the manufacturer will not be liable. Where we are dealing with machine learning proving that the system was defective when it left the hands of the manufacturer will be extremely difficult [12]. This is problematic in the context of the semantic gap in that with autonomous systems, such as in the case of highly automated driving, the operational design domain (operational environment) should be expected to reasonably change over time after the release of the product (for instance the change of behaviour or appearance of other traffic participants), and thus defects may only come to light at a much later stage.

There is also a defence which is included in the statute to encourage technological innovation, which will also protect the manufacturer. It is a defence to prove: "that the state of scientific and technical knowledge at the relevant time was not such that a producer of products of the same description as the product in question might be expected to have discovered the defect if it had existed in his products while they were under his control" [12]. Whilst this defence has been narrowly applied [7], these features of the Consumer Protection Act 1987 mean that claims for injuries resulting from subsequently adapted behaviour may be difficult to sustain. Legally it means that technologies can be deployed even where they are not fully understood.

In addition there is a possible claim against the manufacturer in the tort of negligence where a person suffers actionable harm caused by the manufacturer [61]. Negligence is not strict liability. The injured claimant needs to prove fault on the part of the manufacturer. The manufacturer only owes a duty to the claimant take such care as is reasonable in the circumstances.

This duty is only to do what is reasonable in the circumstances, not to guard against every risk. If a product complies with minimum standards set by government or other relevant organisations, that is generally evidence that the manufacturer has not been negligent. With autonomous systems such standards have not yet been forthcoming, leading to an unclear understanding of what it is reasonable to expect.

Such a claim in negligence will be both difficult and costly. If the system was manufactured in accordance with its design, then the claimant has to prove that the design itself is defective. The claimant also needs to prove that the risk ought to have been foreseen at the time the product left the defendant's control. For instance by demonstrating that the testing of the product was inadequate [32]. In the context of the inevitable semantic gap when dealing with autonomous systems, such a claim will be difficult to sustain.

Further a manufacturer may be justified in taking known or foreseeable risks. If there is a genuine demand for a product, the product is as safe as it can be made at reasonable cost, and the danger posed by the product is not significantly out of proportion to its social benefits, an action in negligence against the manufacturer will fail [29]. Given the significant social benefits of autonomous systems it is likely that the law of negligence will tolerate the semantic gap.

Particularly problematic for a claim in negligence in relation to learned characteristics is the fact that there will be no manufacturer liability in negligence if the product was not dangerous at the time when the manufacturer parted with it, and he had no reason to anticipate that it would become dangerous. Further the complexity and unpredictability of operating domain, makes negligence hard to prove. However, where a product proves particularly dangerous after it has left the defendant's control, there may be a duty to recall it.

The need to prove fault means that negligence claims against manufacturers for learned characteristics will be difficult to prove.

5.2. How the case studies illustrate this gap

The potential for manufacturer liability does not offset the loss of the vicarious liability claim. A significant liability gap has emerged where an autonomous system replaces an employee, to which the semantic gap contributes.

A human driver may commit the tort of negligence, a highly automated driving vehicle may not. In the case of such vehicles the UK Parliament has recognised the need to provide for victims, and has passed the Automated and Electric

Vehicles Act 2018. This Act attempts to solve the liability gap for such vehicles by introducing statutory liability of the vehicle's insurer. This is an *ad hoc* solution to the liability gap and it does not apply outside of this context to other autonomous systems. It also operates in a context where as with other motor vehicles insurance is compulsory – this is one of the very few areas where English law mandates insurance coverage. This statutory solution hides the broader liability gap problem which results from autonomous systems replacing employees, but at least it provides victims with a route of claim where they are injured by such vehicles. Other delivery systems, and other applications of autonomous technologies are not dealt with by the Act and the liability gap remains in these cases.

With the clinical technology case study the system is advisory. A human actor is still present. This means that where the medical practitioner is negligent in conducting the treatment, no liability gap is present since a claim may be brought against the medical practitioner or their employer. However, where it is reasonable to rely on the system, the practitioner will not be negligent, even if the actual decision of the system, or the advice it provides, if it had been generated by a doctor would be negligent. In such a scenario a liability gap is present. But since this gap is only present when the human involved in the medical treatment is not at fault, this liability gap is smaller than if the system operated autonomously. However, it is still present.

The case studies show that the liability gap is greatest for more highly autonomous systems. This is since with an assistive system an individual may in some circumstances still be at fault in implementing the system's recommendations. The key message is any autonomous system, particularly those using machine learning models, will lead to a liability gap when compared to that role being carried out by an employee, unless it is specifically addressed through statutory means. The semantic gap means that this liability gap will not be bridged by the ability to bring claims against manufacturers.

5.3. Suggestions and methods for reducing this gap

The failure to treat autonomous systems as persons for tort purposes limits the liability of users, when compared to those who use employees to do the same work. This creates a tort incentive to replace employees with autonomous systems even where to do so represents a greater risk to third parties. It also encourages a reduction in safety measures.

Tort law deters, particularly at the organisational level [15]. Properly deployed it deters harmful conduct, and encourages safer methods [31]. Where tort encourages the adoption of less safe behaviour this is a perversion of its deterrent role. Tort must be "Tech Neutral"; that is, it should neither encourage nor discourage the adoption of new technologies, where the risks of harm that such technologies pose to third parties are the same when compared to older technologies and methods. Tort should only encourage or discourage adoption where systems are more, or less safe, than the alternatives. With the liability gap tort encourages the adoption of autonomous systems even in situations where they are less safe than their alternatives.

To solve this problem, this article does not advocate that systems should be granted full legal-personhood [8] [54] [59] as this will come with a host of other problems, such as its being able to own property, sue (as well as being sued), and have its own legal rights. In order to attempt to close the liability gap this article suggests a statutory form of vicarious liability for autonomous systems, and that the system should be treated as if it were a person only for the purpose of imposing liability on its user or 'principal'. Autonomous systems can act as agents and form contracts on behalf of their principals [47]. It is not outlandish to suggest that they should be deemed to also be able to act as a form of 'agent' for the purposes of attributing liability to a principal in tort law.

To instead attempt to solve the liability gap merely by strengthening a victim's claims against the manufacturer, means that tort law will not be tech neutral vis-à-vis the system user. Statutory vicarious liability by mimicking the system of liability in place for employees ensures that tort law is tech neutral, and does not encourage or discourage an employer from adopting new technologies to replace employees where their risks of harm are the same. It prevents users of technology escaping liability for their negative externalities by replacing humans with machines. It also helps to close the consequences of the semantic gap. However, care should be taken not to introduce new avenues of claim that would not be present for other technologies or persons. To do so might discourage the adoption of technology, even where it represents a reduced risk for society.

6. A way forward: a multi-disciplinary "gaps" analysis to inform safety assurance

This paper has discussed the relationship between the semantic, responsibility and liability gaps in autonomous systems. We have illustrated these gaps using examples from the highly automated driving and clinical advisory domains. These gaps are currently hindering the development and adoption of technologies that have the potential for huge societal benefits. A summary of the semantic, responsibility and liability gaps is given in Table 1.

6.1. The way forward: immediate steps

An analysis of the gaps across design, ethics, and law can contribute to ethically- and legally-informed engineering processes, and engineering-informed ethics and legal practice. In turn, this offers a promising approach for grounding the justification of safety requirements for autonomous systems.

Table 1
Summary of the semantic, responsibility and liability gaps.

	Semantic gap	Responsibility gap	Liability gap
Description	Normal conditions for complete specification are not present	Normal conditions for moral responsibility are not present	Normal conditions for civil liability are not present
Root causes	Complexity and unpredictability of the operating environment	Complexity and unpredictability of the operating environment	Complexity and unpredictability of the operating environment
	Complexity and unpredictability of the system	Complexity and unpredictability of the system	Complexity and unpredictability of the system
	Transfer of decision function to the system	Transfer of decision function to the system	Transfer of decision function to the system
When incurred	Design phase	All phases	All phases
Responsible parties	Manufacturer	Manufacturer, operator, and user	Manufacturer, operator, and user
Risk	Acceptable safety is under-specified	No one is morally accountable for harm	Loss falls on the victim

The integration of these multi-disciplinary perspectives could be achieved by addressing the following three "touch points" between the disciplines, and this can be done based on current understanding of the technical, ethical and legal issues, and the descriptions of the gaps we introduced above.

Definition of the operational domain: A better understanding of the ethical and legal expectations on the system for a given operational domain, including tolerable residual risk, known laws and precedents, and different cultural attitudes, contributes to a more holistic definition, which enables a more complete specification of the system. A good definition of the operational domain can give a sound basis for reducing the semantic gap, and for contributing towards the understanding and managing of responsibility and liability gaps. Hereby, a deep understanding of the technical limitations of the situation shall be used to evaluate the actual risk caused by the system for any given context.

Detailed system design and evaluation: Safety-critical scenarios should be identified during the system design phase. We have suggested the method of reflective equilibrium to determine technically feasible, ethically justified system behaviour in different types of scenarios. A legal representative can also be included in this reflective process in the design phase. Simulation techniques and field validation data in particular would allow for variations of scenarios to be generated and evaluated in order better to define the acceptable boundaries of the system behaviour. If an acceptable response cannot be achieved, then a restriction of the operational design domain or an alternative system design may be required.

Post-hoc analysis and allocation of blame and liability: In liability cases, an equally balanced cross-disciplinary process could be applied to analyse the specific incident in detail, based on the available information and multiple perspectives included in the initial reflective design process.

The idea is not that these multi-disciplinary touch points are applied sequentially across the development process, but rather that the interactions occur continually to ensure that the semantic gap is iteratively minimised. Future work should develop the suggestions detailed above and apply them to real-world case studies in order to derive a set of concrete recommendations for future development, standardisation and regulation.

6.2. The way forward: emerging approaches

There are some additional novel, or emerging, approaches which we believe would help to address these issues – to reduce the gaps. What is proposed here is largely aspirational – although there is ongoing work developing these approaches:

1. **Safety Case** – a key part of the safety case should define how gaps are identified and what has been done to control them. For example, the argument should include a demonstration of the delimitation of moral and legal responsibility between designer/manufacturer, operator and the user. Fig. 3 outlines the safety assurance dimensions relevant to the semantic gap. Similar models are needed for the other gaps, and for the detailed problems of the use AI and machine learning in the system; see [4] for ideas on these issues. Acceptance of the safety case should be based on the method of reflective equilibrium, achieving an acceptable level of coherence amongst the considered judgements and beliefs of representatives from as many relevant stakeholder groups as possible.
2. **Monitoring and Dynamic Assurance** – due to the complexity and uncertainty of the system and its operational environment, it is likely that the initial safety case will not accurately reflect the actual gaps – be they semantic, responsibility or legal. The systems of interest are data-rich, and this gives the opportunity to assess what is actually happening, as opposed to what was anticipated at the design stage. We can use machine learning to process the operational data, to determine whether or not the initial safety case was accurate, and to identify the need for remedial action if not [37], with some initial validation of the ideas in the healthcare domain (treatment of atrial fibrillation following thoracic surgery). This work may ultimately enable the safety case to be updated dynamically to reflect actual risks and gaps, although this is most likely to be practicable for the semantic gap.

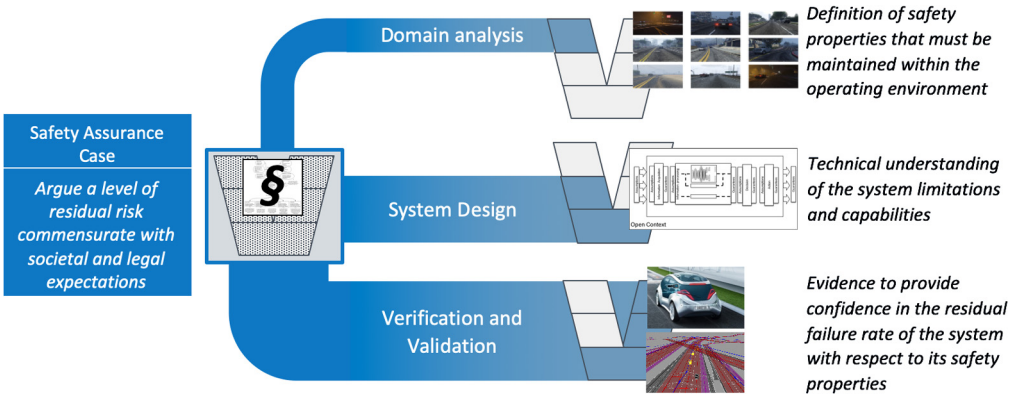

Fig. 3. Safety Assurance Dimensions.

3. **Soft Law** – legal and regulatory processes change slowly, and it is difficult to keep pace with technological change. Whilst some changes in the 'hard law' might be desirable or essential, e.g. to produce 'tech neutral' tort law, other changes which we refer to as 'soft law' might help mitigate some of the legal gaps. An example might be an industry-wide code of practice, which companies adopt. The code of practice could also include provisions on the compensation of those harmed by autonomous systems, which would help to address part of the liability gap arising from the semantic gap. If companies do not comply with the code, then this could lead to qualification of annual reports when the company is audited. Soft law may also help to shape the standards expected by the courts of actors in the industry, for instance when courts determine claims in negligence. Another advantage of soft law is that it may be capable of influence across different jurisdictions. Further, standards bodies are aware of the need to be more agile, and some are producing Publicly Available Specifications (PAS) or other pre-normative standards as this can be done rapidly; the IEEE P7000 series, are a good example of this.
4. **Regulation and Governance** – the characteristics of AI and autonomous systems, especially where used in critical situations, challenge traditional approaches to regulation and governance; it will be hard for regulatory frameworks to keep up with the pace of change, and governance will also need to be increasingly agile. Whilst 'soft law' approaches might help here, there are real challenges in regulating and governing such technologies. Points that need to be considered include; explicit balancing of benefits against risk; the notion of 'contestability' of decisions (this may be more practicable than explainability), and accountability. There have been many publications on principles for 'ethical AI', including [57]; whilst there are some similarities in the published principles, many are expressed rather more as 'values', e.g. harmony, than principles that can be realised. There remains a significant issue in implementing regulatory and governance principles, given the pace of change and the geographic spread of the technologies, the scope of supply chains, etc. Encouragingly, in the UK the government is funding work on making regulation more adaptive, including supporting "regulatory sandpits", which are analogous with "safe harbours", that enable experimentation with agile regulatory principles.

6.3. Closing remarks

In the London underground and in other transportation systems travellers are exhorted to 'mind the gap' when boarding or alighting from trains, where the gap between the train and the platform is seen as a source of risk. It seems to us that this might also be a useful exhortation for those developing, deploying, and regulating AI and autonomous systems. Our hope is that by finding systematic ways of identifying and categorising 'gaps' we can assist in exposing the sources of risk with AI and autonomous systems, and thus contribute to risk control. This paper has shown that the notion can be used in technical, ethical and legal contexts and we believe that it might be a unifying concept, and one that can also be applied in more detailed models of assurance [37].

Declaration of competing interest

We declare no conflict of interest in submitting the paper titled "Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective".

Acknowledgements

This work was partially supported by the Assuring Autonomy International Programme (https://www.york.ac.uk/assuring-autonomy).

References

[1] Dario Amodei, et al., Concrete problems in AI safety, arXiv preprint, arXiv:1606.06565, 2016.
[2] J.M. Anderson, K. Nidhi, K.D. Stanley, P. Sorensen, C. Samaras, O.A. Oluwatola, Autonomous Vehicle Technology: A Guide for Policymakers, Rand Corporation, 2014 Jan 10.
[3] Aristotle, Nichomachean Ethics, Terence Irwin (Ed.), Hackett, 1985.
[4] R. Ashmore, R. Calinescu, C. Paterson, Assuring the machine learning lifecycle: desiderata, methods, and challenges, arXiv preprint, arXiv:1905.04223, 2019.
[5] Attorney-General of the British Virgin Islands v. Hartwell [2004] UKPC 12; Mattis v. Pollock [2003] EWCA Civ 887; [2003] 1 WLR 2158.
[6] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.F. Bonnefon, I. Rahwan, The moral machine experiment, Nature 563 (7729) (2018 Nov) 59.
[7] A v National Blood Authority [2001] 3 All ER 289.
[8] Susanne Beck, The problem of ascribing legal responsibility in the case of robotics, AI Soc. 31 (2016) 473.
[9] Carl Bergenhem, et al., How to reach complete safety requirement refinement for autonomous vehicles, in: CARS 2015-Critical Automotive Applications: Robustness & Safety, 2015.
[10] M. Coeckelbergh, Responsibility and the moral phenomenology of using self-driving cars, Appl. Artif. Intell. 30 (8) (2016) 748–757.
[11] Council Directive 85/374/EEC.
[12] Paulius Cerka, Jurgita Grigiene, Gintare Sirbikyt, Liability for damages caused by artificial intelligence, Comput. Law Secur. Rev. 31 (2015) 376, 386.
[13] Norman Daniels, Reflective equilibrium, in: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), https://plato.stanford.edu/archives/fall2018/entries/reflective-equilibrium/.
[14] Armen Der Kiureghian, Ove Ditlevsen, Aleatory or epistemic? Does it matter?, Struct. Saf. 31 (2) (2009) 105–112.
[15] Don Dewees, David Duff, Michael Trebilcock, Exploring the Domain of Accident Law, OUP, 1996;
Peter Cane, Atiyah's Accidents Compensation and the Law, 8th edn., CUP, 2013, pp. 421–452, see also, 9th edn., CUP, 2018, p. 422.
[16] Federal Ministry of Transport and Digital Infrastructure, Ethics commission on automated driving, https://www.bmvi.de/SharedDocs/EN/PressRelease/2017/084-ethic-commission-report-automated-driving.html, August 2017.
[17] C. Fleischmann, A. Scherag, N.K.J. Adhikari, C.S. Hartog, T. Tsaganos, P. Schlattmann, et al., Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations, Am. J. Respir. Crit. Care Med. 193 (3) (2015 Sep 28) 259–272.
[18] P. Feth, R. Adler, T. Fukuda, T. Ishigooka, S. Otsuka, D. Schneider, D. Uecker, K. Yoshimura, Multi-aspect safety engineering for highly automated driving, in: International Conference on Computer Safety, Reliability, and Security, Springer, Cham, 2018, pp. 59–72.
[19] J.M. Fischer, M. Ravizza, Responsibility and Control: A Theory of Moral Responsibility, Cambridge University Press, 1998.
[20] Philippa Foot, The Problem of Abortion and the Doctrine of Double Effect, 1967.
[21] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016.
[22] J.E. Gotts, M.A. Matthay, Sepsis: pathophysiology and clinical management, BMJ 353 (2016 May 23) i1585.
[23] David Gunkel, The Machine Question: Critical Perspectives on AI, Robots, and Ethics, MIT Press, 2012.
[24] R. Hakli, P. Mäkelä, Moral responsibility of robots and hybrid agents, Monist 102 (2019) 259–275.
[25] Brody Huval, et al., An empirical evaluation of deep learning on highway driving, arXiv preprint, arXiv:1504.01716, 2015.
[26] IEEE, IEEE Standard Adoption of ISO/IEC 15026-1 - Systems and Software Engineering - Systems and Software Assurance, IEEE, New York, USA, 2014.
[27] ISO, ISO 26262: Road Vehicles - Functional Safety, ISO, Geneva, Switzerland, 2011.
[28] ISO, ISO/PRF PAS 21448: Road Vehicles - Safety of the Intended Functionality, ISO, Geneva, Switzerland, 2018.
[29] Michael Jones (Ed.), Clerk & Lindsell on Torts, 22nd edn., Sweet and Maxwell, 2018, pp. 11–29.
[30] M. Komorowski, L.A. Celi, O. Badawi, A.C. Gordon, A.A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nat. Med. 24 (11) (2018 Nov) 1716.
[31] William Landes, Richard Posner, The Economic Structure of Tort Law, Harv UP, 1987.
[32] Law Commission, Automated vehicles, https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2018/11/6.5066_LC_AV-Consultation-Paper-5-November_061118_WEB-1.pdf, November 2018.
[33] M.M. Levy, L.E. Evans, A. Rhodes, The surviving sepsis campaign bundle: 2018 update, Intensive Care Med. 44 (6) (2018 Jun) 925–928.
[34] P. Lin, Why ethics matters for autonomous cars, in: M. Maurer, J.C. Gerdes, B. Lenz, H. Winner (Eds.), Autonomes fahren: Technische, rechtliche und gesellschaftliche aspekte, Springer, Berlin, Heidelberg, 2015, pp. 69–85.
[35] Andreas Matthias, The responsibility gap: ascribing responsibility for the actions of learning automata, Ethics Inf. Technol. 6 (3) (2004) 175–183.
[36] John McDermid, Self-driving cars: why we can't expect them to be moral, The Conversation (January 2019).
[37] J. McDermid, J. Jia, I. Habli, Towards a framework for safety assurance of autonomous systems, in: Workshop on Artificial Intelligence Safety 2019, Macao, China, August, 2019.
[38] D. McPherson, C. Griffiths, M. Williams, A. Baker, E. Klodawski, B. Jacobson, et al., Sepsis-associated mortality in England: an analysis of multiple cause of death data from 2001 to 2010, BMJ Open. 3 (8) (2013 Aug) e002586.
[39] J. Millar, Ethics settings for autonomous vehicles, in: Patrick Lin, et al. (Eds.), Robot Ethics 2:0, 2017.
[40] Phillip Morgan, Vicarious liability on the move, Law Q. Rev. 129 (2013) 139, approved in Allen & Ors v The Chief Constable of the Hampshire Constabulary [2013] EWCA Civ 967.
[41] Jonathan Morgan, Torts and technology, in: Roger Brownsword, Eloise Scotford, Karen Yeung (Eds.), The Oxford Handbook of Law, Regulation and Technology, OUP, 2017, ch. 22.
[42] Merel Noorman, Deborah G. Johnson, Negotiating autonomy and responsibility in military robots, Ethics Inf. Technol. 16 (1) (2014) 51–62.
[43] NTSB, Office of Public Affairs, News Release, Preliminary Report Released for Crash Involving Pedestrian, Uber Technologies, Inc., Test Vehicle, https://www.ntsb.gov/news/press-releases/pages/nr20180524.aspx, 24 May 2018.
[44] NTSB, Executive Summary, Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, https://www.ntsb.gov/investigations/accidentreports/pages/har1702.aspx, May 7, 2016.
[45] S. Nyholm, J. Smids, The ethics of accident-algorithms for self-driving cars: an applied trolley problem?, Ethical Theory Moral Pract. 9 (5) (2016) 1275–1289.
[46] S. Nyholm, Attributing agency to automated systems: reflections on human-robot collaborations and responsibility-loci, Sci. Eng. Ethics 24 (2018) 1201–1219.
[47] Ugo Pagallo, The Laws of Robots, Crimes, Contracts, and Torts, Springer, 2013, pp. 95–102 (at least in Anglo-American law, but not in some other legal systems).
[48] John Rawls, Outline of a decision procedure for ethics, Philos. Rev. 60 (2) (1951) 177–197.
[49] J. Rawls, A Theory of Justice, Harvard University Press, 1971.
[50] SAE, SAE J3016 Surface Vehicle Recommended Practice, (R) Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, SAE, 2018.
[51] F. Santoni de Sio, J. van den Hoven, Meaningful human control over autonomous systems: a philosophical account, Front. Robot. AI (28 February 2018), https://doi.org/10.3389/frobt.2018.00015.

[52] W.H. Self, M.W. Semler, R. Bellomo, S.M. Brown, B.P. deBoisblanc, M.C. Exline, et al., Liberal versus restrictive intravenous fluid therapy for early septic shock: rationale for a randomized trial, Ann. Emerg. Med. 72 (4) (2018 Oct) 457–466.
[53] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, et al., The third international consensus definitions for sepsis and septic shock (sepsis-3), JAMA 315 (8) (2016 Feb 23) 801–810.
[54] S.M. Solaima, Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy, Artif. Intell. Law 25 (2017) 155.
[55] Rob Sparrow, Killer robots, J. Appl. Philos. 24 (1) (2007) 62–77.
[56] Judith Jarvis Thomson, The trolley problem, Yale Law J. 94 (1984) 1395.
[57] UK RAS Network, Ethical issues for robotics and autonomous systems, https://www.ukras.org/wp-content/uploads/2019/07/UK_RAS_AI_ethics_web_72.pdf, 2019.
[58] US Supreme Court's decision in Wyeth v. Levine, 555 U.S. 555 (2009), although note the pending appeal of Merck Sharp and Dohme Corp. v. Albrecht. No. 17-290.
[59] Robert van den Hoven van Genderen, Do we need new legal personhood in the age of robots and AI?, in: Robotics, AI and the Future of Law, Springer, Singapore, 2018, pp. 15–55.
[60] Gerhard Wagner, Comparative tort law, in: Mathias Reiman, Reinhard Zimmermann (Eds.), The Oxford Handbook of Comparative Law, OUP, 2006, pp. 1005–1010.
[61] Christopher Walton (Ed.), Charlesworth & Percy on Negligence, 14th edn., Sweet and Maxwell, 2018, pp. 16–78, 16–118.
[62] Waymo, Waymo safety report - on the road to fully self-driving, www.daimler.com/innovation/case/autonomous/safety-first-for-automated-driving-2.html, 2017.
[63] L. Yilmaz, A. Franco-Watkins, T. Kroecker, Coherence-driven reflective equilibrium model of ethical decision-making, in: IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2016.
[64] Konrad Zweigert, Hein Kötz, Introduction to Comparative Law, 3rd edn, OUP, 1998, Tony Weir tr.