|  |  |  |  |
| --- | --- | --- | --- |
| № | Оригинал (En) | Перевод (Ru) | Комментарии проверяющего |
| 1 | Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective | Остерегайтесь пробелов: обеспечение безопасности автономных систем с инженерной, этической и правовой точек зрения |  |
| 2 | 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. | АННОТАЦИЯ  Данная статья объединяет междисциплинарный подход с точки зрения системной инженерии, этики и права для формулировки общего языка, на котором можно рассуждать о многогранной проблеме обеспечения безопасности автономных систем. |  |
| 3 | 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. | Основное внимание в работе уделяется «пробелам», возникающим в процессе разработки: семантическому пробелу, когда отсутствуют нормальные условия для полной спецификации предполагаемой функциональности; пробелу в обязанности, когда нет нормальных условий для поддержания в людях моральной ответственности за ущерб; и пробелу в ответственности, когда отсутствуют нормальные условия для обеспечения компенсации жертвам вреда. |  |
| 4 | By categorising these “gaps” we can expose with greater precision key sources of uncertainty and risk with autonomous systems. | Классифицируя эти «пробелы», мы можем с большей точностью выявить ключевые источники неопределенностей и рисков для автономных систем. |  |
| 5 | This can inform the development of more detailed models of safety assurance and contribute to more effective risk control. | Это может помочь в разработке более подробных моделей обеспечения безопасности и способствовать более эффективному контролю рисков. |  |
| 6 | 1. Introduction  Autonomous systems disrupt established practices of system design, moral responsibility, legal liability, and safety assurance. | 1. Введение  Автономные системы нарушают сложившуюся практику проектирования систем, моральной ответственности, юридической ответственности и обеспечения безопасности. |  |
| 7 | By “autonomous system,” we mean a system that makes decisions independently of human control. | Под «автономной системой» мы понимаем систему, которая принимает решения независимо от контроля человека. |  |
| 8 | Autonomous decision-making capability can be understood as a spectrum. | Возможность автономного принятия решений можно представить в виде спектра. |  |
| 9 | 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. | В верхней части этого спектра находятся высоко автономные системы, которым мы делегируем как процесс принятия решений, так и реализацию последующих действий. |  |
| 10 | 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. | В нижней части – консультативные системы, которым мы делегируем некоторую или большую часть процесса принятия решений, но выполнение рекомендованных действий является обязанностью ответственного человека. |  |
| 11 | This paper considers the spectrum, but illustrates the ideas by considering both types of autonomous system in case studies. | Эта статья учитывает спектр, но иллюстрирует идеи, рассматривая оба типа автономной системы в тематических исследованиях. |  |
| 12 | The disruption that autonomous systems present can be explained in terms of ‘gaps’ that arise across the design and implementation process. | Нарушения в работе автономных систем можно объяснить с помощью «пробелов», возникающих в процессе проектирования и реализации. |  |
| 13 | We look at three gaps: the semantic gap; the responsibility gap; the liability gap. | Рассмотрим три пробела: семантический пробел, пробел в обязанности, пробел в ответственности. |  |
| 14 | The semantic gap arises because the normal conditions are not met for manufacturers, particularly designers, to provide a complete specification of the system. | Семантический пробел возникает из-за того, что нормальные условия для предоставления полной спецификации системы не соблюдаются для производителей, в особенности для проектировщиков. |  |
| 15 | The semantic gap represents the difference between the implicit intentions on the system’s functionality and the explicit, concrete specification that is used to build the system. | Семантический пробел представляет собой разницу между неявными намерениями по функциональности системы и явной, конкретной спецификацией, которая используется для построения системы. |  |
| 16 | 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. | Риск данного пробела возникает на этапе проектирования, но это также постоянная проблема, требующая внесения изменений в спецификацию после внедрения и развертывания. |  |
| 17 | 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. |  |  |
| 18 | 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. |  |  |
| 19 | 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. |  |  |
| 20 | 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. |  |  |
| 21 | 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. |  |  |
| 22 | These three issues also affect the safety assurance of autonomous systems. |  |  |
| 23 | 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. |  |  |
| 24 | 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. |  |  |
| 25 | We ask the reader to note some clarifications. |  |  |
| 26 | First, we place particular emphasis on autonomous systems that use machine-learning models rather than rule-based models. |  |  |
| 27 | Second, for brevity, we refer to three kinds of human actor: the manufacturer, the operator, and the user. |  |  |
| 28 | The term “manufacturer” should be understood as having a wide scope, including designers, developers, programmers, engineers, manufacturing companies. |  |  |
| 29 | Third, this paper is primarily concerned with accident and injury as a result of an autonomous system’s decision. |  |  |
| 30 | 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. |  |  |
| 31 | The paper is structured as follows. In Section 2, we define and analyse the semantic gap, and its implications for safety assurance. |  |  |
| 32 | In Section 3, we illustrate the semantic gap with two case studies: a highly automated driving system and a clinical advisory system. |  |  |
| 33 | 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. |  |  |
| 34 | In section 5, we consider the liability gap. |  |  |
| 35 | 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. |  |  |
| 36 | In Section 6, we build on this multi-disciplinary analysis to propose an approach to the safety assurance of autonomous systems. |  |  |
| 37 | 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]. |  |  |
| 38 | There are three sources of the Semantic gap. We refer to these as the three root causes throughout the paper. |  |  |
| 39 | The first root cause is the complexity and unpredictability of the operational domain. |  |  |
| 40 | Autonomous systems typically operate within an environment that cannot be fully specified at design time. |  |  |
| 41 | This is due to inherent environmental complexity at any given point in time, as well as continual, ongoing changes to the domain. |  |  |
| 42 | Together these constitute an “open context,” for which a complete specification is very difficult, if not impossible, to formalise. |  |  |
| 43 | The second root cause is the complexity and unpredictability of the system itself. |  |  |
| 44 | Because of the computational techniques used, the system is inherently complex. |  |  |
| 45 | It will also change continually through interactions within the domain. |  |  |
| 46 | 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. |  |  |
| 47 | In addition, the system has technical limitations, such as its incomplete “understanding” of the environment, which restricts its ability to react to certain situations. |  |  |
| 48 | All of these factors arising from the complexity of the system make a complete specification problematic. |  |  |
| 49 | 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). |  |  |
| 50 | This entails that new functions are introduced to the systems, including those that historically required human interpretation, ethical judgement and lawful behaviour. |  |  |
| 51 | 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. |  |  |
| 52 | This further complicates the process of specifying intended functionality. |  |  |
| 53 | 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. |  |  |
| 54 | Various methods are typically used to confirm that the system specification meets user needs. |  |  |
| 55 | These methods range from the systematic capture, analysis, and review of requirements to statistically representative field-based tests. |  |  |
| 56 | But this is far more difficult against a backdrop of environmental complexity and unpredictability, along with increasing decision-making being transferred to the system. |  |  |
| 57 | Here, safe system performance cannot always be guaranteed. |  |  |
| 58 | 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. |  |  |
| 59 | Fig. 1 summarises the functional components of a highly automated driving system. |  |  |
| 60 | 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). |  |  |
| 61 | 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). |  |  |
| 62 | 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. |  |  |
| 63 | Action components execute the driving strategy via the set of vehicle actuators (i.e. brakes, engine, steering column, etc.). |  |  |
| 64 | 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. |  |  |
| 65 | Deep neural networks can make sense of unstructured data using efficient computations in real-time [21]. |  |  |
| 66 | 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]. |  |  |
| 67 | At present, two strands of development exist for highly automated driving. |  |  |
| 68 | The first is SAE Level 3: Conditional Driving Automation Systems. |  |  |
| 69 | These take over control of the vehicle whilst driving on highways [50]. |  |  |
| 70 | 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. |  |  |
| 71 | These systems are an evolution of current driver assistance systems. |  |  |
| 72 | There is therefore a degree of familiarity with their capabilities and their technical limitations. |  |  |
| 73 | The second application of automated driving is SAE Level 4: “High Driving Automation” systems for urban automated driving. |  |  |
| 74 | 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. |  |  |
| 75 | This is a more revolutionary approach to mobility. |  |  |
| 76 | It may take the form of not just classical passenger cars but also new classes of vehicles, such as driverless shuttles or delivery vehicles. |  |  |
| 77 | These vehicles will travel at significantly lower speeds than with SAE Level 3, but their technical and safety assurance challenges are far greater. |  |  |
| 78 | 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. |  |  |
| 79 | SAE Level 4 in particular shows how the three root causes described in Section 2 contribute to the Semantic gap. |  |  |
| 80 | 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. |  |  |
| 81 | 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. |  |  |
| 82 | It is therefore impossible to create a complete specification of the operating environment for urban automated driving. |  |  |
| 83 | Second, the systems themselves need to be more complex to manage the driving task in urban environments. |  |  |
| 84 | Heterogeneous sensing channels are required to counteract the individual weaknesses of each sensor type (e.g. camera, radar etc.). |  |  |
| 85 | 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. |  |  |
| 86 | 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. |  |  |
| 87 | 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. |  |  |
| 88 | They seem particularly well-suited to learning functionality that cannot be easily specified using traditional procedural means (if X happens, then do Y). |  |  |
| 89 | But there is a catch. |  |  |
| 90 | Machine learning functions do not deliver clear-cut answers. |  |  |
| 91 | 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. |  |  |
| 92 | In addition, the processes which lead to these decisions are difficult to decipher. |  |  |
| 93 | These attributes result in a paradox or “no free lunch” effect, where the problem of deriving a suitable specification of the intended behaviouris instead transferred to the problem of demonstrating that the implemented (learned) behaviour meets the intent. |  |  |
| 94 | The third root cause that contributes to the semantic gap with SAE Level 4 systems is the significantly increased transfer of decision function. |  |  |
| 95 | 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. |  |  |
| 96 | This is a huge task, involving considerable uncertainty. |  |  |
| 97 | 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. |  |  |
| 98 | *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. |  |  |
| 99 | This results in a different class of safety requirements to those previously considered in the industry [50]. |  |  |
| 100 | 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. |  |  |
| 101 | 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. |  |  |
| 102 | The conditions for acceptable functional safety for passenger vehicles are set by the international functional safety standard for road vehicles ISO 26262 [27]. |  |  |
| 103 | This standard is limited to hazards, i.e. sources of harm, caused by the vehicles’ malfunctioning behaviour. |  |  |
| 104 | This remains necessary to ensure that the hardware is reliable and the implementation of systems is fault tolerant. |  |  |
| 105 | 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]. |  |  |
| 106 | As a result, additional approaches must be developed. |  |  |
| 107 | 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]. |  |  |
| 108 | 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. |  |  |
| 109 | It must justify the acceptable level of residual risk1 associated with this function.  (1The risk that remains once all risk reduction measures have been taken.) |  |  |
| 110 | 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]. |  |  |
| 111 | The first fatal accidents caused by vehicles operating in this driving mode have highlighted the need to discuss and reach con-sensus on acceptable residual risk for such systems [43][44]. |  |  |
| 112 | This wider discussion should also include the potential of the systems to significantly increase road safety [2][18]. |  |  |
| 113 | *3.1.4. Suggestions for reducing the semantic gap* |  |  |
| 114 | A number of approaches are currently being considered to close the semantic gap during the design of highly automated cars. |  |  |
| 115 | These approaches target the gap’s three root causes. |  |  |
| 116 | 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. |  |  |
| 117 | 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 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  (2The 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.) |  |  |
|  | 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 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 Section2relate 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 rep-resentation 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. |  |  |
|  | t 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*. |  |  |