UNRAVELLING RESPONSIBILITY FOR AI

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

ABSTRACT

To reason about where responsibility does and should lie in complex situations involving AI-enabled systems, we first need a sufficiently clear and detailed cross-disciplinary vocabulary for talking about responsibility. Responsibility is a triadic relation involving an actor, an occurrence, and a way of being responsible. As part of a conscious effort towards 'unravelling' the concept of responsibility to support practical reasoning about responsibility for AI, this paper takes the three-part formulation, 'Actor A is responsible for Occurrence O' and identifies valid combinations of subcategories of A, is responsible for, and O. These valid combinations – which we term "responsibility strings" – are grouped into four senses of responsibility: role-responsibility; causal responsibility; legal liability-responsibility; and moral responsibility. They are illustrated with two running examples, one involving a healthcare AI-based system and another the fatal collision of an AV with a pedestrian in Tempe, Arizona in 2018. The output of the paper is 81 responsibility strings. The aim is that these strings provide the vocabulary for people across disciplines to be clear and specific about the different ways that different actors are responsible for different occurrences within a complex event for which responsibility is sought, allowing for precise and targeted interdisciplinary normative deliberations.

INTRODUCTION

Systems with machine learnt (ML) components, and particularly those which operate with limited direct and continuous human input, raise difficult questions for responsibility.[2] These systems can often interpret data and make 'decisions' which can directly affect human wellbeing, but they do not currently possess the capacities that warrant ascriptions of legal or moral responsibility for the outcomes they cause (Thoma, 2021; Burton et al., 2020). The systems themselves cannot be legally or morally responsible. But the normal criteria or conditions for ascribing forms of liability and moral responsibility to natural and legal persons are also not clearly met for the outcomes these systems cause. The 'root causes' of this are: the transfer of decision-making functions which historically have been carried out by human operators; the fact that the intended functionality of the systems cannot be completely and explicitly specified; the inherent complex, uncertainty and opacity of many ML models; the systems' capacities for adaptive performance; and the unpredictability of events in the operating environment. In the presence of these factors, robust human control and foresight is constrained, and causation difficult to prove. As such, responsibility and liability 'gaps' can arise (Morgan, 2024; Morgan, 2023; Burton et al., 2020; Matthias, 2004).

Another source of difficulty for ascribing responsibility for AI is the Problem of Many Hands (Thompson, 1980). This is the problem that there is a proliferation of actors involved in a causal chain

[1] Corresponding author: zoe.porter@york.ac.uk
The authors would like to thank Paul Noordhof for his comments on a draft of this paper.

[2] We use the term 'AI-enabled system' to denote a hardware system with AI components, and specifically, ML-components (viewing ML as a subcategory of AI). We take it that AI-enabled systems can have varying degrees of independence from direct human intervention.

of events leading to complex outcomes, which complicates and apparently precludes ascriptions of individual responsibility for these outcomes. An existing problem in many multi-actor networks, including public administration (Pesch, 2008) and R&D networks (Zwart et al., 2006), it also arises with the deployment of advanced, software-intensive systems (Cooper et al., 2022; Nissenbaum, 1996). Engineers, for example, often install components into hardware systems which depend on software prepared by others, such as commercial-off-the-shelf components (NIST/CNSSI 4009-2015); and no single person is responsible for all software development decisions (Nissenbaum, 1996). In addition, the ML pipeline in particular is multi-stage, with different groups designing, training, evaluating, deploying and monitoring ML models, and ML engineers often relying on toolkits built by others (e.g., libraries of algorithm implementations, model architectures, and large-scale, pre-trained models) (Cooper et al., 2022). In the presence of these factors, it is difficult, and perhaps impossible, to discern members of a set of individuals responsible for the outcomes of AI-enabled systems; indeed, some argue that such cases can only be made sense of in terms of collective responsibility (Van de Poel, Royakkers & Zwart, 2015).[3]

This paper addresses the following problem: in order to reason about where responsibility does and should lie in complex cases and situations involving AI-enabled systems, we first need a sufficiently clear and detailed cross-disciplinary vocabulary for talking about responsibility. Achieving this is non-trivial because 'responsibility' is a rich concept, with multiple overlapping meanings both in our everyday talk and across disciplines. We adopt the metaphor of 'unravelling responsibility' to describe our approach. The work is not an attempt to provide answers to normative questions of where responsibility does or should lie for particular occurrences involving AI. Rather, the scope of the paper is to provide a vocabulary, which is accessible to all stakeholders in the debate, to articulate those questions in a clear and specific way.

The paper is structured as follows:

- *Section 2* presents two scenarios which are used to illustrate the discussion in Sections 3 and 4. In both these scenarios, a complex chain of decisions, actions and omissions, involving an AI-enabled system, lead to a harmful consequence. The first scenario is hypothetical. It concerns an ML-based decision support system used to predict diabetes comorbidities. The second scenario is the fatal collision of a vehicle with an Automated Driving System (ADS) with a pedestrian in Tempe, Arizona in 2018.
- *Section 3* sets out the framework for the vocabulary proposed. It introduces the three-part formulation 'Actor *A* is responsible for Occurrence *O*', which functions as the central structuring device for 'unravelling responsibility'. Each of the three parts – *A*, is responsible for, and *O* — is broken down into subcategories.
- *Section 4* identifies 'responsibility strings'. Each responsibility string is a valid combination of the subcategories of *A*, is responsible for, and *O*. The strings are grouped into four clusters,

[3] 'Collective responsibility' is generally taken to refer to the responsibility, and more specifically the moral responsibility, of a group as an irreducible entity. 'Shared responsibility' by contrast is more individualistic, referring to the moral responsibility of discrete individuals who acted together to bring about an outcome (Smiley, 2022). This discussion is picked up below in Section 3.2 and Section 4.4. To note, in the original formulation of the Problem of Many Hands (Thompson, 1980), Thompson argued that deferring to collective responsibility was not the correct response to the problem and that, with careful application of individualist criteria for moral responsibility, individual responsibility could be found more often than is assumed.

according to four senses of responsibility: role responsibility; causal responsibility; legal liability responsibility; moral responsibility. These are described in more detail and conditions for each are outlined.

- *Section 5* considers how the approach taken to 'unravelling responsibility' could be useful to the multidisciplinary debate about responsibility for AI. It concludes with two next steps for the work. First, a systematic presentation of the strings in a visual map or graph which could be used to pick out different "responsibility pathways" between actors and occurrences, and to reason normatively about the location of responsibility for occurrences involving AI. Second, the adoption of the vocabulary set out in this paper to inform and structure responsibility assurance cases.[4]
- The Appendix contains and enumerates the full set of unravelled responsibility strings.

2. EXAMPLES: DIABETES AI AND UBER TEMPE

To start, we introduce two running examples which will be used to illustrate the discussion. For shorthand, the first example is referred to as 'Diabetes AI' and the second example as 'Uber Tempe'. These scenarios involve multiple actors, including AI-enabled systems, and depict a harmful outcome (potential in the case of the Diabetes AI).

2.1 Diabetes AI

'Diabetes AI' is a system which uses an ensemble of different Machine Learning (ML) models, including Naive Bayes (NB), Neural Network (NN), Random Forest (RF) and Support Vector Machine (SVM) to support clinical decision-making in the treatment of patients with diabetes Type II (Ozturk et al., 2023; Ryan Conmy et al., 2023a). Its intended use is to predict a patient's risk of developing a diabetes comorbidity, such as high blood pressure, or having a catastrophic event, such as heart attack, within the next six months.

Diabetes AI is trained on real (anonymised) data records of patients with diabetes Type II. These records document hundreds of patient observations, such as Body Mass Index (BMI), platelet and red blood cell counts, and creatinine levels. As with most medical AI systems which use real patient data, the raw dataset distribution suffers a number of issues which can introduce bias. There is over-representation of observations from patients who have had Type II diabetes for a long time; there are also missing results (e.g. historical BMI scores) and under-represented patient groups (e.g. patients of ethnicities with a higher risk of certain comorbidities). There are several mitigation methods such as data imputation which can be used to mitigate missing results. In addition, using methods such as k-fold cross-validation during model training can reduce overfitting (which occurs when the model performs very well on the training dataset but poorly in the real world). But exhaustive testing of the model is precluded by the extensive variability of patients, comorbidities and progression of the condition. The ML methods that are used to mitigate training data issues can also introduce new uncertainty. Training data bias therefore remains an issue that can affect the accuracy of predictions for some patients.

[4] When, in this paper, we use the elliptical phrase 'responsibility for AI', the intended, more precise meaning is 'responsibility for occurrences involving AI'.

Here is a hypothetical scenario involving Diabetes AI in which the location of responsibility for various occurrences may be sought. A patient, who was diagnosed with Type II diabetes last year, attends a six-monthly appointment at the diabetes clinic run by the local NHS Trust. The focus of the consultation is the patient's risk of developing high blood pressure, and subsequent risk of heart attack. During the consultation, the clinician's direct observations and examinations lead to the judgement that the patient is possibly at high risk of developing dangerously high blood pressure. The prediction from Diabetes AI is that the patient is low-risk. Unbeknown to clinician and patient, this is a false negative prediction from the AI. It is due to over-representation in the training data of patients who have lived with diabetes for a long time, which the patient has not. In addition, during the consultation, the patient slightly exaggerates how much they have been making lifestyle changes (i.e. to diet and exercise). On the basis of these three informational inputs (direct patient observations, Diabetes AI's low-risk prediction and the patient's report), the clinician decides the patient's blood pressure does not point to risk of imminent heart attack. The clinician therefore does not prescribe medication to control hypertension, and strongly advises the patient to keep up with the reported lifestyle changes. The patient fails to do so. Six weeks later, after some days of feeling unwell, the patient has a heart attack.

2.2 Uber Tempe

The AI-based system in the 'Uber Tempe' scenario was an adapted Volvo SUV (Sports Utility Vehicle), which incorporated an Automated Driving System (ADS) and was developed and operated by the Uber Advanced Technologies Group (ATG). When activated, the ADS had full control of the vehicle. The ADS used three systems to detect and classify objects in the environment – Lidar, Radar and image-based – and fused them to create and update predicted paths based on object type and trajectory. The image system was also used to recognise traffic signs and lights. In the test vehicle depicted in the scenario below, the ADS could only be activated in pre-mapped areas. The vehicle operator (safety driver) could disengage the ADS. The ADS could also deactivate and hand over to the safety driver if it encountered a situation it could not handle, which it might do with warning or suddenly.

The Uber Tempe scenario concerns a real-world human fatality. On 18 March 2018 at 21:58, on a four-lane section of highway in Tempe, Arizona, the test vehicle collided with a pedestrian, Elaine Herzberg, pushing a bicycle across the road, causing her fatal injuries. The accident report identified a number of issues which contributed to the crash (NTSB, 2019). There were several immediately preceding causal factors; each of these contributed to the fact that the collision was not prevented. The safety driver, tasked with monitoring the road and intervening to prevent collisions and unsafe scenarios, did not intervene in time. The safety driver was not looking at the road. In addition, the time it took for the ADS to determine a collision was imminent meant the system did not provide an audible warning to the safety driver until about a second before collision.[5] Furthermore, an automated braking system in the SUV – which would have prevented the collision – was disabled by Uber ATG employees to avoid possible radar signal interference. Wider contributory factors to the

[5] Although the ADS detected a moving object, it was unable to classify consistently that this was a pedestrian/bicycle and instead thrashed between different classifications. Each time a new classification was determined prior information on the object's trajectory was deleted, making it impossible to predict its path until the AV was very close, hence the delay in warning the safety driver.

collision were also identified in the accident report, including: insufficient safety monitoring during, and risk mitigation systems in, the overall design of the AV; ineffective monitoring by Uber ATG of the performance of safety drivers; the policy of removing a second safety driver from the test vehicles; a lack of "safety culture" at Uber ATG; limited regulatory requirements from Arizona State and the Transportation Safety Board.

3. ACTOR *A* IS RESPONSIBLE FOR OCCURRENCE *O*

This section sets out the framework for the vocabulary proposed. Responsibility describes a kind of relation between an actor and an occurrence (Shoemaker, 2015). Responsibility is ascribed *to* actors. Actors are responsible *for* occurrences. As such, the central structuring device and basis for 'unravelling' responsibility is the abstract, three-part formulation: Actor *A* is responsible for Occurrence *O*. This provides for identifying the 'responsibility strings': valid combinations of subcategories of *A*, is responsible for, and *O*.

3.1. The relevant senses of 'is responsible for'

Responsibility concerns the ways in which actors can be answerable for occurrences. This reflects the etymology of the word 'responsible', which comes from the Latin *respondere*, meaning 'to answer'.

The first step in unravelling responsibility is to distinguish the key senses of responsibility, or the main kinds of ways that actors can be answerable for occurrences. We adapt H.L.A. Hart's enduring taxonomy of responsibility (Hart, 1968/2008, p. 210-227),[6] as shown in Table 1 below:

Sense of responsibility	Description
role-responsibility	*A* has duties that attach to their role
causal responsibility	*A* is a cause of *O*
legal liability-responsibility	*A* is liable to legal sanction/redress for *O*[7]
moral responsibility	*A* is an author of *O* (attributability) *A* is liable to moral sanction for *O* (accountability)

Table 1: The main senses of 'is responsible for'[8]

[6] Hart's taxonomy of responsibility is the starting point for many introductions to the concept of responsibility (see, for example, Zerilli et al., 2021; Kutz, 2004). Modern taxonomies of (specifically moral) responsibility, which reference Hart, can be found in: Vincent, 2011; Van de Poel 2015.

[7] This sense is about *legal liability*, not *legal responsibility* more generally (a duty to act in accordance with the law). Legal liability, insofar as it is concerned with outcomes, is concerned exclusively with *adverse* outcomes. By contrast, there could be a positive consequence of legal responsibility, e.g., fulfilling a contractual duty may lead to a right to payment.

[8] A further key sense of responsibility which Hart includes (1968/2008, p.227-230) is *capacity-responsibility*: the possession of psychological capacities, such as understanding and reasoning, which is a criterion for people being morally responsible and, in most cases, liable for what they do. Capacity-responsibility is a condition of some forms of role-responsibility, most cases of legal liability and all forms of moral responsibility. In the

Using the Diabetes AI responsibility scenario to illustrate:

- *Role-responsibility*: the clinician has a well-defined role-responsibility, namely professional duties to the patient. Other actors, such as the developers, the manufacturers, NHS Trust managers and regulatory officials will also have explicit and implicit duties that attach to their roles.
- *Causal responsibility:* several actors were causes of, or made causal contributions to, the heart attack, including (and not limited to) two omissions: the clinician's failure to prescribe medication to manage hypertension; the patient's failure to make lifestyle changes that would minimise the risk of heart attack.
- *Legal liability-responsibility*: if we imagine that the heart attack was caused by the clinician's medical negligence in virtue of falling below the standard of a reasonable clinician in being unduly influenced by a false negative AI prediction, this would mean the clinician had committed a tort (a civil wrong) and could be liable.[9]
- *Moral responsibility*: it seems several actors (including the clinician, the patient and the developers of Diabetes AI, given the false negative recommendation) *potentially* bear some moral responsibility (in the sense of attributability) in virtue of contributing to the outcome through their voluntary conduct, although the degree to which they could control or foresee the heart attack is both variable and unclear. Were any of them morally responsible (in the sense of attributability), the appropriateness of further holding them morally responsible (i.e. morally accountable) would need to be determined by reference to questions such as whether their conduct had fallen below the standard expected of them and whether holding them responsible would improve future practice.

3.2 Actor *A*

The second step is to delineate the kinds of actors who may be responsible for occurrences. To avoid the baggage attached to the word 'agent', the term 'actor' is adopted instead, defined very generally as an entity that does something and can cause change in the world. Actors are the logical subjects of responsibility ascriptions. In any given responsibility scenario involving AI, there will be multiple actors involved at various points in the lifecycle and across the wider socio-technical system.

Three main subcategories of 'actor *A*' delineated in this paper are, using the Uber Tempe example: 1. AI-enabled system (e.g. the adapted SUV); 2. Individual human (e.g. the individual pedestrian, safety driver, engineer, manager or official); 3. Institution (e.g. Uber ATG).[10]

---

example instantiations of the responsibility strings for those types of responsibility in Section 4 below, it is assumed to obtain in those cases.

[9] The reasonable standard of care is for a clinician using an AI-enabled decision support system is still unclear. For discussions of the liability of clinicians using AI, see: Lawton et al., 2023; Jones, Thornton & Wyatt, 2023; Tobia, Nielsen & Stremitzer, 2021.

[10] An institution, which is typically incorporated, may take a number of legal forms, including (amongst others) a private or public limited company, or corporate sole. More rarely an institution may be an unincorporated association which lacks legal personality. This paper does not seek to deal with unincorporated associations which are in law, merely a collection of individuals bound together by contract. As a reducible group, this category is covered by individual human actors.

To clarify, 'AI-enabled system' here refers just to the machine, and not to the wider system that includes the human using the machine. While development of the work may include a subcategory of 'human-AI team', presently these could be constructed by combining AI-enabled systems and individual human actors. Furthermore, for simplicity, we do not include a subcategory of *human collectives*. As the work develops, 'human collective' may be added as a further subcategory of 'actor *A*'. There is a distinction between human collectives understood as groups which transcend their individual members and human collectives understood as groups which reduce to their individual members. We may call the former 'conglomerate collectives' and the latter 'aggregate collectives' (French, 1984). For now, it should be taken that aggregate collectives could be constructed from individual actors (picked out by the second subcategory). Conglomerate collectives are to a large extent picked out by the subcategory 'institution'.

While the term 'actor' is universal, not all actors are equal in terms of the kinds of responsibility that can be ascribed to them. An AI-enabled system, such as an SUV with an activated ADS, is an actor in a "thin" sense. It does things and causes change in its environment, but it is not a moral agent. Moral agents have the ability to act voluntarily, or to choose to act, with reference to their understanding of right and wrong. Only moral agents can be morally responsible for what they do. We adopt the mainstream view that AI-enabled systems are not moral agents.[11] An AI is also not a legal person. Only legal persons can owe legal duties or enter into contracts, and hence be liable for civil law wrongs or be subject to criminal prosecution. In not being a legal person, an AI-enabled system cannot be legally liable for harm or damage it causes.[12] Individual humans (natural persons) are, by contrast, both moral agents and legal persons. Institutions are legal persons and, though this is disputed, we assume they are not moral agents.[13]

To summarise, the kinds of actor A delineated in this paper are:

- AI-enabled system (which is neither a moral agent nor a legal person);
- Individual human (who is both a moral agent and a legal person);
- Institution (which is a legal person but not a moral agent).

3.3 Occurrence *O*

The third step is to categorise the kinds of 'occurrence *O*' for which responsibility may be sought. These are broken down as follows. One group concerns *the AI-enabled system's* outputs. These are

[11] By most accounts, contemporary and near-future forms of AI-enabled systems fall short of the agential capacity required to warrant an ascription of moral responsibility (Verdicchio and Perin, 2022; Véliz, 2021; Coeckelbergh, 2020; Johnson, 2006). For the view that they can be moral agents but not morally responsible, see: Floridi and Sanders, 2004; Wallach and Allen, 2009.

[12] For discussions on AI and legal personhood, see: Chesterman, 2020; Turner, 2019; Pagallo, 2013.

[13] It is a point of philosophical debate whether conglomerate collectives can be moral agents and hence appropriate bearers of moral obligation and moral responsibility. Those who argue against the notion of conglomerate collectives having moral agency (and the viability of collective, as opposed to shared, responsibility) do so either on the basis of methodological individualism, which challenges the very notion of irreducible group moral agency (because, for example, such entities cannot have intentions), or on the basis of normative individualism, which holds that collective responsibility is unfair (Smiley, 2022). For opposing views, see: French, 1998; Tuomela, 1989; Pettit, 2007; List and Pettit, 2011. It is beyond the scope of the present work to go into greater detail on this live debate in philosophy.

its internal state classifications, predictions or recommendations (denoted 'decision*'), its external transmission or implementation of these, in physical actions and manoeuvres (denoted 'action*'), or its failure to take an action (denoted 'omission*'). The second group concerns the *individual human's* or the *institution's* 'outputs'. These are called 'decisions', 'actions' and 'omissions' (without the star). In the individual human case, we may think of a 'decision' as a conclusion reached after consideration or deliberation, an 'action' as a thing done intentionally, and an 'omission' as an action not performed. In the institutional case, we may think of a 'decision' as a conclusion reached through organisational procedures, and 'action' as things done by the collective or on behalf of the institution, and 'omissions' as things not done.[14] The third group is *consequences*. Consequences are the outcomes that are caused by *any* of the other subcategories of *O*, as well as by physical events in the operating environment.

The main reason for distinguishing the system's outputs from human and institutional decisions and actions is to help those reasoning about responsibility in complex cases involving AI to distinguish between a cause of an occurrence from an actor who cannot be liable or morally responsible and one who can. It should be noted that the distinction both between a 'decision*' and an 'action*', as well as the distinction between a 'decision' and an 'action', is somewhat blurred, and this will be considered in the development of the work.

Table 2 gives individual examples of each subcategory of O, using Diabetes AI:

Decision*	Diabetes AI's (false negative) prediction
Action*	Diabetes AI's passing on its decision* to a wider display system
Omission*	There is no example of an omission* in the scenario as described, but if Diabetes AI had failed to execute an action*, it would be an omission*
Decision	The clinician decides the patient's blood pressure does not place him at risk of imminent heart attack
Action	The clinician examines the patient
Omission	The clinician does not prescribe medication
Consequence	The patient has a heart attack

Table 2: Subcategories of *O*

4. UNRAVELLING THE RESPONSIBILITY STRINGS

This section identifies what we call the "responsibility strings". The aim is that these will enable people to be clear and specific about the different ways that different actors are responsible for different occurrences within a complex event for which responsibility is sought. Each valid

[14] An additional relevant distinction is the difference between decision/action/omission *types or classes*, and *specific or token* decisions/action/omissions. This distinction and its incorporation into the framework will be explored in the development of the work.

combination of the subcategories of *A*, is responsible for, and *O* is referred to as an unravelled 'responsibility string': a distinct way in which '*A* is responsible for *O*' can be understood. Not all combinations of subcategories are valid. For example, strings of '*A* is morally responsible for *O*' where A is an AI-enabled system would not be valid, because of the stipulation that AI-enabled systems cannot bear moral responsibility. Further exceptions are identified in the relevant subsections below.

Here, in Section 4, the responsibility strings are unravelled. It should be noted that a further decomposition of role-responsibility, legal liability-responsibility and moral responsibility is given. In total, 85 responsibility strings are identified. The overall model of these is presented in the Unified Modelling Language (UML) in Figure 1 below. This model does not cover exceptions or, in other words, it does not restrict permitted combinations. The full enumeration of the valid strings is presented in the Appendix.

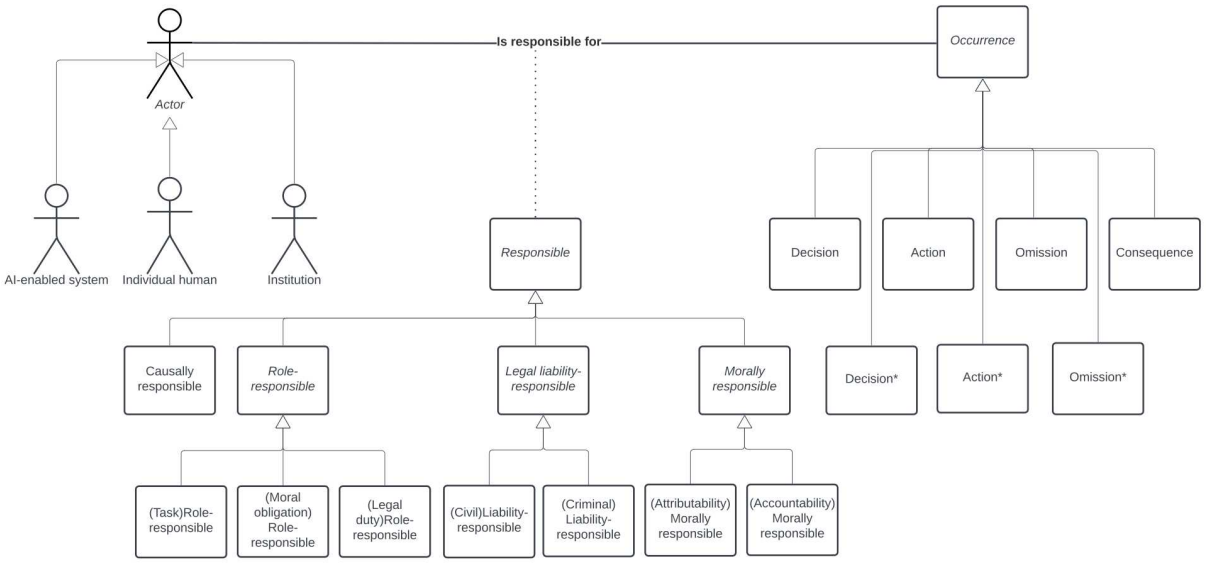

Figure 1

The scope of this paper is limited to 'unravelling' the distinct ways in which 'A is responsible for O' can be understood. It provides the building blocks for later connecting responsibility strings, but making these systematic connections is a next step for the work, as discussed in Section 5.

4.1. *A* is role-responsible for *O*

People often talk about their or other people's responsibilities. When they do so, they generally mean the tasks, duties, and obligations that attach to various roles within society or organisations. As Hart puts it, *"...whenever a person occupies a distinctive place or office in a social organization, to which specific duties are attached to provide for the welfare of others or to advance in some specific way the aims or purposes of the organization, he is properly said to be responsible for the performance of these duties"* (Hart, 2008, p. 212). Following Hart's characterisation, we refer to this type of responsibility as 'role-responsibility'.

Being a 'responsible person' or 'behaving responsibly' is associated with role-responsibility (Hart, 2008).[15] Role-responsibility is also the sense of 'responsibility' in Responsible Research and Innovation policy (Owen et al., 2013) and Responsible AI (Peters et al., 2020; Schiff et al., 2020; Dignum, 2019; Askell, Brundage & Hadfield, 2019), which concern actors taking seriously their professional and moral duties to design, engineer, manufacture, use and govern innovative technologies in a way that steers them towards desirable societal goals (Von Schomberg, 2013).

Before unravelling the role-responsibility strings, we subdivide 'role-responsibility' into three further categories: tasks; moral obligations; and legal duties. The same activity might sometimes fall into more than one of these categories simultaneously, and there may also be conflicts between them. Moreover, despite delineating them clearly for the purposes of this paper, it should be noted that the boundaries of one's role responsibilities are not always sharply defined.

Role-dependent tasks are those tasks, functions and duties that are assigned to *A*'s role in an organisation or group. All categories of *A* may have role-dependent tasks. *Role-dependent moral obligations* are what morality requires of us. Broadly speaking, we have moral obligations not to harm others unjustly (because they have a right not to be harmed or wronged) and sometimes moral obligations to come to the aid of others.[16] Only moral agents (i.e., individual humans, on this paper's stipulation), can have moral obligations. *Role-dependent legal duties* are set out in law. An actor may have a duty of care (i.e., a duty not to cause actionable harm by failing to meet an objective standard of care set by the standard of a reasonable person carrying out the same function) (Charlesworth & Percy, 2018). The law may also prescribe other standards, for instance where a party has contracted to achieve a particular result, or where *A*'s activity is unusually hazardous. In some cases, a legal person may have an absolute duty (i.e., a duty that must be adhered to whatever the effort, time, and cost) towards those exposed to the risk of their activities. Only legal persons (i.e., individual humans and institutions) can have legal duties.

On this decomposition of role-responsibility, there are 42 role-responsibility strings. Table 3, in the Appendix, enumerates each role-responsibility string individually. Before illustrating a selection of these strings below, some clarifications should be made. First, the strings are not exclusive. It is possible that more than one actor is role-responsible for the same occurrence, the same actor will often be role-responsible for more than one occurrence, and some kinds of actor may be role-responsible in more than one way for an occurrence (e.g., it might be an ML engineer's task, moral obligation and legal duty to mitigate bias in a training dataset). Second, these strings will in practice often be 'chained' (e.g., in being role-responsible for performing an action, one might be role-responsible for making several decisions). Third, there are connections between string-types (e.g., in fulfilling a role-responsibility for *O* one would also be causally responsible for *O*, although the converse does not hold, in that one might cause something it was not one's role to cause), but the scope of the present work is simply to unravel the discrete possibilities as a precursor to future work in which they are systematically connected.

Uber Tempe can be used to illustrate a selection of role-responsibility strings. The numbers correspond to those in Table 3 (in the Appendix). To be clear, these are intended as illustrative

[15] Some accounts classify this as responsibility as a virtue (Vallor, 2023; Van de Poel, Royakkers & Zwart, 2015).
[16] Doctors, for example, have a moral obligation to come to the aid of others in an emergency, even when off-duty.

examples; in the example below (R19), we are not picking out a specific engineer, for example, and identifying a moral obligation they actually held; rather, we are using this as a plausible example of what such an obligation could be.

	A	**Is role-responsible for**	***O***
R2.	AI-enabled system	Is (task)role-responsible for	Action*

E.g., R2. ADS is (task)role-responsible for executing the dynamic driving task whilst activated.

R19.	Individual human actor	Is (moral obligation)role-responsible for	Action

E.g., R19. System engineer is (moral obligation)role-responsible for conducting system safety analysis.

R26.	Individual human actor	Is (legal duty)role-responsible for	Action

E.g., R26. System engineer is (legal duty)role-responsible for taking reasonable care (assessed at the standard of the reasonable AV systems engineer) to avoid acts or omissions which they can reasonably foresee would be likely to cause injury to another.[17]

R35.	Institution	Is (task)role-responsible for	Consequence

E.g., R33. Uber ATG is (task)role-responsible for the safety and performance of its ADS.

R40.	Institution	Is (legal duty)role-responsible for	Action

E.g., R40. Uber ATG is (legal duty)role-responsible for following safety standards.

4.2. *A* is causally responsible for *O*

'Causal responsibility' is another way of referring to causation. To say *A* was causally responsible for *O* is just to say *A* caused *O*. This does not mean that *A* was the *only* cause of *O*, or even the *most important* cause of *O*. It simply means that *A* was *a* cause of *O*.

This paper does not divide causal responsibility into further subcategories. According to an 'egalitarian' notion of cause, any component of a set of causal circumstances is properly called a cause, including enabling conditions, causal factors, and omissions (Noordhof, 2020, p. 177). On stricter notions of cause, something has to have a special feature, such as being either a triggering condition or an unusual intrusion in the normal state of affairs (e.g., the poison in the tea rather than the drinking of it) to count as a cause (Noordhof, 2020, p. 185-187). For simplicity, we assume here an egalitarian understanding of causal responsibility.

To note, however, we are focused only on the causal responsibility of the three types of *actors* delineated (AI-enabled system, individual human and institution), and not on the conditions and

[17] A duty of care is not freestanding; duties are owed to particular persons, in relation to a particular form of harm.

physical events that might also be said to be causally responsible for outcomes. Moreover, for enquiries into liability and moral responsibility, the ascription of causal responsibility is selective: *"what we wish to do is select from the welter of causal factors that made some contribution to the event in question one to be denominated the cause"* (Feinberg, 1970, p. 142). For such enquiries, therefore, this egalitarian notion of causation will need to be supplemented by other considerations.

On the present approach to unravelling causal responsibility, there are 21 causal responsibility strings. Because any actor can cause any kind of occurrence, causal relations between all *A*s and *O*s are valid. Table 4 in the Appendix enumerates each causal responsibility string individually. Before illustrating a selection of these strings below, similar clarifications to those with respect to role-responsibility should be made. First, the strings are not exclusive. It is possible, indeed likely, that more than one actor is causally responsible for the same occurrence. For example, in the example C4 below, the AI-enabled system (in virtue of its decision* that the patient is low-risk) is not the only cause of the clinician's final decision that the patient is low-risk. Second, being causally responsible for *O* does not require being the immediate causal antecedent of *O*. Third, and connected to this, these strings will often also be 'chained'. For example, in the example C8 below, the ML engineer would be causally responsible for intermediate decisions and actions (e.g., about the dataset and model testing) which determine the system's decisions*. As before, individual "strings" provide the building blocks to create these chains where it is helpful to do so.

	A	**Is causally responsible for**	***O***
C2.	AI-enabled system	Is causally responsible for	action*

E.g., C1. Diabetes AI is causally responsible for displaying its prediction to the clinician (and patient).

C4.	AI-enabled system	Is causally responsible for	decision

E.g., C4. Diabetes AI is causally responsible for the clinician's final decision that the patient is low-risk.

C8.	Individual human	Is causally responsible for	decision*

E.g., C8. ML engineer is causally responsible for Diabetes AI's prediction that the patient is low-risk.

C13.	Individual human	Is causally responsible for	omission

E.g., C13. Patient is causally responsible for not implementing lifestyle choices.[18]

C19.	Institution	Is causally responsible for	action

E.g., C19. NHS Trust is causally responsible for installing Diabetes AI in five clinics in the area.

[18] A central debate in the philosophy of action is between agent causation (agents cause events) and event causation (events cause events). Proponents of event causation, which is the more standard position, deny that causation by agents can be understood in any way other than by event causation (i.e., it is in virtue of those events which are the person's mental states that their actions and omissions are caused). Not much turns on this for the purposes of the present inquiry. More important to note is that C13 is not yet to say that the patient is morally responsible for the omission (if it could be shown that it was not voluntary on the patient's part, then the patient would be causally but not morally responsible for it).

4.3. *A* is legal liability-responsible for *O*

'Legal responsibility' describes an obligation to act in accordance with the law and it stems from a legal duty (see section 4.1), which may be attached to a role or assumed, for instance by agreement. Legal responsibility is a necessary but insufficient condition of liability. The focus in this cluster of responsibility strings is liability. As Hart puts it, *"When legal rules require men to act or abstain from action, one who breaks the law is usually liable, according to other legal rules, to punishment for his misdeeds, or to make compensation to persons injured thereby"* (Hart, 1967/2008, p. 215).[19]

While liability regimes vary from jurisdiction to jurisdiction, the following is broadly true across all legal systems. Criminal law concerns the prosecution of behaviour, which has been elevated to the status of a criminal offence, in the criminal courts (most often by the state). It primarily aims to safeguard the public, punish harmful behaviour and deter misconduct. It is concerned with (amongst other things) the protection of bodily integrity, the protection of property, and of the public welfare, and as such, criminalises offences against the person, and criminalises certain culpable violations of property rights, and so on.[20] Civil law (sometimes called private law – and not to be confused with Civil law jurisdictions, a term given to legal jurisdictions that have their roots in Roman law) regulates behaviour between parties. It seeks to determine the rights and duties of legal persons, for example by establishing civil liability for a harm or a wrong. The same conduct might concern more than one category of law (e.g., it might be both a criminal offence and a tort – a civil wrong). Public law regulates the behaviour of public bodies. It is concerned with the rights, powers, and duties of the state. For present purposes, public law is omitted from the analysis.

This paper divides legal liability-responsibility into criminal and civil liability only. Whilst the analysis focuses on the common law,[21] these principles are found across developed legal systems, many having a common taxonomic or historical root.

A criminal offence generally requires an action element (*actus reus*) and a mental element (*mens rea*). The action element refers to the person's conduct and/or the consequences of their actions (or omissions). The mental element refers to the person's 'state of mind' at the time of the offence and explores questions of intent, recklessness, and knowledge of their conduct. Both elements must coincide in time (e.g. a person cannot refer to their state of mind at a later point in time) *and* must correspond to the alleged offence. Criminal law embodies a wide range of criminal offences. Two kinds of criminal liability may prove salient to legal cases involving AI-enabled systems. First, strict liability, which is often used in regulatory offences and is sometimes referred to as no-fault liability. This lacks the *mens rea* element – there is thus no required state of mind such as intention or recklessness. Second, secondary liability, which considers those who aided or encouraged the

[19] Noting that, particularly in civil law, actors can sometimes be liable for the actions of other actors (i.e., vicarious liability), as discussed in this subsection.

[20] Criminal law goes beyond the criminalisation of direct offences against the person to also include the protection of public goods and public welfare (e.g., rules for public transport and road use).

[21] This includes England and Wales, the United States, Australia, Ireland, New Zealand, Canada, India, Hong Kong, and Singapore, (amongst others).

completion of a criminal offence by a principal offender. If the offence was that of strict liability, the principal's *mens rea* is not required – but it is required for the accessory to prove their intention.[22]

Unlike criminal proceedings, civil liability claims arise between private parties. Two areas of civil liability are particularly pertinent to cases involving AI-enabled systems. First, tort law, (called delict in some, primarily civilian jurisdictions, such as France) which covers negligence, product liability, mechanisms of secondary tortious liability, vicarious liability, and statutory direct actions against insurers (amongst others) – all of which are influenced by the existence of insurance. Second, contract law, which covers obligations, liability and damages between parties who enter into binding agreements and form contractual relationships. It may also include insurance contracts. For brevity, liability claims in contract law arise in cases where a party breaches the terms of an agreement and fails to fulfil their contractual obligations. An in-depth discussion of contract law is beyond the scope of this section.

Different torts have different elements. The workhorse tort in common law jurisdictions is negligence. Its core elements are:[23] the defendant (i.e., the legal person who actioned the harm) owes the claimant (i.e., the injured party) a duty of care in law, the duty is breached by the defendant not reaching the required standard of care applicable, and the breach caused the claimant's harm. A further element explores legal causation. Some jurisdictions treat this as part of the causation analysis (cause in law), whereas others treat it as an additional remoteness of damage element.

Product liability is concerned with harm or damage caused by defective products. Modern product liability law evolved in the United States, before influencing other jurisdictions. Within the European Union the rules are laid out in the Product Liability Directive.[24] This area of tort covers manufacturers' liability and compensatory measures for injured parties. This is distinguished from contractual liability, although some areas may overlap. *Vicarious liability* (often called *respondeat superior* in a United States context) is a mechanism in tort law whereby a legal person can be held liable for a tort committed by another person. Such mechanisms are common across many legal systems, having a long historical tradition (Giliker, 2018). It is a form of secondary liability. Similarly, *secondary tortious liability* considers situations where a legal person is treated as a joint wrongdoer and is subsequently held liable for harmful actions or omissions committed by the primary tortfeasor. Lastly, in *statutory actions against insurers*, a claim can arise between the injured party and the defendant's liability insurer. Examples include the European Union's Fourth Motor Insurance Directive and the UK's Automated and Electric Vehicles Act 2018, which attaches liability for damages to the insurer[25] when accidents are caused by an automated vehicle.

---

[22] Criminal liability is an especially nuanced area. It is difficult to concisely capture the various ways in which a person is held liable for their actions without discussing questions of capacity (i.e., the right and ability of a person to make particular decisions, have legal duties, enter into legal contracts, and so forth), causation, and foreseeability, to name a few. This brief discussion, therefore, has attempted only to provide a basic distillation of the concepts in criminal law.

[23] As previously noted, the law of tort also considers other areas where liability for harm arises and remedies are awarded.

[24] This approach is also implemented into UK law via the Consumer Protection Act 1987 which remains in force post-Brexit.

[25] Automated and Electric Vehicles Act 2018, s 2(1).

On a decomposition of legal liability-responsibility into criminal and civil liability, the approach taken in this paper is to unravel 4 legal liability-responsibility strings. Because only legal persons are eligible for legal liability, *A* is limited to individual humans and institutions. Because of liability's inherently backwards-looking nature, and its typical requirement that an adverse consequence has actually occurred, the only *O*s are consequences.[26] Table 5, in the Appendix, enumerates each liability string. To note, there will be connections between strings identifying legal duties and causal responsibility with strings identifying liability; however, to repeat, the purpose here is to set out the distinct strings as a precursor to connecting them in a development of the work.

Diabetes AI can illustrate the liability strings. In the examples below, the scenario has been varied to make appropriate instantiations of the strings. The *O* in each case is the patient's heart attack; let us imagine that it is a fatal heart attack.

	A	**Is legal liability-responsible for**	***O***
L1.	Individual Human	Is (criminal)liability-responsible for	Consequence

E.g., L1. If the clinician had deliberately prescribed incorrect medication with the intent of precipitating a fatal heart attack in the patient, the clinician would be criminally liable for the fatality.

L2.	Individual Human	Is (civil)liability-responsible for	Consequence

E.g., L2. If it were demonstrated that the fatal heart attack was directly attributable to the clinician's negligence in failing to reject Diabetes AI's incorrect recommendation (i.e., falling below the standard of care expected of the reasonable clinician), the clinician could be civilly liable through tort law.

L3.	Institution	Is (criminal)liability-responsible for	Consequence

E.g., L3. If it were demonstrated that a failing in the running of the NHS Trust by senior management, such as failures to provide adequate support and training for staff, was a cause of and a substantial element in the death by heart attack in the patient, then, under the Corporate Manslaughter and Corporate Homicide Act 2007, the NHS Trust could be charged with corporate manslaughter and be criminally liable for the consequence.

L4.	Institution	Is (civil)liability-responsible for	Consequence

E.g., L4. To pick up on example L2. above, the NHS Trust could be vicariously liable for the negligence of its employee, the clinician, under civil law.

4.4 *A* is morally responsible for *O*

'Moral responsibility' means A is answerable for an *O* in a way that opens A up to moral appraisal and, possibly, sanction from the wider community. In everyday speech, one often hears references

[26] Some forms of liability are triggered merely by the breach of duty and do not require a particular harm. For now, such cases of liability have not been included in the decomposition.

"to people's "moral responsibility" where the point is to indicate that a person has some duty or obligation—some responsibility—to which that person is required, by some standard, to attend" (Talbert, 2022). These are picked out by the (moral obligation)role-responsibility strings in the first group. Here, we are concerned with the relation an *A* bears to their actions or omissions, and their consequences, in order for it to be appropriate to hold *A* accountable for them.

One simple and perhaps useful distinction is that between *being morally responsible* and *being held morally responsible*. Following Watson's distinction, we refer to this as 'moral responsibility as attributability' and 'moral responsibility as accountability', respectively (Watson, 1996).

First, there is being morally responsible, or *moral responsibility as attributability*. For the purposes of this paper, we can say A is responsible in this sense for *O* when A has voluntarily performed or voluntarily caused *O*.[27] In a nutshell, this is the distinction between causal responsibility for *O* and moral responsibility as attributability for *O*: the latter involves voluntary agency. Taking an Aristotelian approach (Aristotle, 2002), this means that one was the source of *O*, did not cause *O* in ignorance, nor under extreme pressure or duress.[28] Moral responsibility as attributability is about *"the relation of an individual to her behaviour"* (Watson, 1996, p. 229). Another way of putting it is that the individual is the author of her behaviour. In the vocabulary of the responsibility strings, this behaviour is broken down into decisions, actions and omissions. Being morally responsible (as attributability) for consequences generally turns on whether it was reasonably foreseeable that these consequences would be caused by the individual's behaviour.

Second, there is being held morally responsible, or *moral responsibility as accountability*. A moral agent *A* is morally accountable when she is liable to a range of reactions and responses, including sanction, from the social forum or moral community. Moral accountability concerns not just the relation of the individual to her behaviour but also her *interpersonal relations*, and involves individuals or groups holding each other *"to certain expectations or demands or requirements"* (Watson, 1996, p. 235). Here, a connection with moral obligations can be made (see section 4.1), since one way a person fails to meet expectations is by failing to fulfil their moral obligations. Though moral responsibility does not solely concern blame but also encompasses praise and a wide range of positive reactive attitudes (Strawson, 1962), such as gratitude and admiration, it does *often* concern blame, in practice and within the literature. One who bears moral responsibility as attributability for *O* might arouse judgements of blameworthiness. When these feelings are expressed and involve sanctions of some kind, such as punishment or demands for public apology, or involve *"dispositions to treat others in generally unwelcome ways"* (Watson, 1996, p. 238), then *A* is morally accountable.

It is a basic assumption amongst philosophers that being morally responsible is necessary for being justly held morally responsible (Watson, 2004, p. 263, 278). In other words, there is a presumed entailment relationship between accountability and attributability (Shoemaker, 2013, p. 160). Sometimes, an *A* might be attributionally responsible but not morally accountable for *O* because

[27] This is not Watson's way of describing responsibility as attributability. For Watson, an actor is attributionally responsible when their behaviour expresses their values and character. The approach taken in this paper, that an actor is attributionally responsible for *O* when *O* is the product of their voluntary agency, derives from Aristotle (2002).

[28] 'Causing O in ignorance' would include not being aware of what one was doing or not being aware of, or not reasonably being able to foresee, the consequences of one's action.

people might consider robust blaming responses to be inappropriate or unfair in the circumstances. *A* might not have egregiously failed to fulfil a moral obligation, for example. And sometimes holding people morally accountable would not lead to better future consequences. This is at the root of 'just safety culture' theories (Dekker and Breakey, 2016), where the emphasis is on 'no blame' disclosures of mistakes and learning from experience. Other times, an A who *should* be held morally accountable might not be. One concern in the wider computing and AI ethics landscape is that people will hide in the complexities of technology to avoid sanction and censure by the moral community (Nissenbaum, 1996; Ananny and Crawford, 2016; Rubel, Castro & Pham, 2019; Porter et al., 2022).

Because only moral agents can bear moral responsibility, the only *A*s who can instantiate moral responsibility strings, on this paper's stipulations, are individual human actors. On this decomposition of moral responsibility, there are 18 moral responsibility strings. The full set of moral responsibility strings are enumerated in Table 6 in the Appendix. Again, to note, the strings are not exclusive: in many real-world cases involving AI, multiple *A*s will likely instantiate the same strings; moral responsibility for an *O* will often not be borne by a single actor alone (these would be cases of shared moral responsibility). As before, there are also connections between strings. For example, causal responsibility is a necessary (but insufficient) condition of moral responsibility; as such, in describing a scenario in which moral responsibility is sought or identified, moral responsibility strings would be connected to causal responsibility strings. However, as already noted, the approach taken in this paper is simply to unravel or decompose the possibilities before connecting them in the development of work.

We can now illustrate the moral responsibility strings, using the Uber Tempe example. In the examples below, the scenario has been varied to make appropriate instantiations of the strings. To be clear, these are intended as illustrative examples; in most of the examples below, we are not picking out a specific, a known senior manager (M9; M12) or a known senior regulatory official (M13); rather, we are using these as speculative examples, since the accident report does not identify individuals (NTSB, 2019).

	A	**Is morally responsible for**	***O***
M7.	Individual human	Is (attributability)morally responsible for	Decision

E.g., M7. Safety driver is (attributability)morally responsible for her decision to watch videos on her phone while the ADS was activated.

M8.	Individual human	Is (accountability)morally responsible for	Decision

E.g., M8. Safety driver is (accountability)morally responsible for her decision to watch videos on her phone while the ADS was activated.

M9.	Individual human	Is (attributability)morally responsible for	Action

E.g., M9. Senior manager of Uber ATG is (attributability)morally responsible for the removal of the back-up safety driver in every car.

M12.	Individual human	Is (accountability)morally responsible for	Omission

E.g., M12. Senior manager of Uber ATG is (accountability)morally responsible for failing to put in place an adequate safety culture.

M13.	Individual human	Is (attributability)morally responsible for	Consequence

E.g., M13. Senior regulatory official is (attributability)morally responsible for human fatalities as a reasonably foreseeable consequence of the choice to allow AVs to be tested on the roads without oversight.

5. Discussion

This paper has 'unravelled' the concept of responsibility to provide a clear, detailed vocabulary to support interdisciplinary reasoning about where responsibility does, or should, lie for occurrences involving AI. The output is a set of 85 unravelled 'responsibility strings', which are enumerated in Tables 3-6 in the Appendix below. Its scope is limited to providing some underpinning conceptual clarity and specificity around the different ways that actors can be responsible for different occurrences within a complex event for which responsibility is sought. The paper has not attempted to answer difficult normative questions about, for example, what various actors' tasks, obligations and duties in the lifecycle should be, nor what liability regimes would be most effective to regulate the development and use of AI, nor who should be held morally accountable for adverse outcomes. It is nonetheless hoped that the vocabulary provided will be of use to such discussions, and particularly those discussions which integrate the perspectives of engineering, law and philosophy, (two disciplines with a long history of analysing responsibility) to reach fitting normative conclusions.

As an illustration of some of the ways the approach may be useful, consider the following. The *role-responsibility* strings, set out in terms of tasks, moral obligation and legal duties, can provide a basis for identifying conflicts between an actor's tasks, moral obligations or legal duties. The *causal responsibility* strings can help us to avoid an easy assimilation of causal responsibility and moral responsibility for occurrences. They could also be used to articulate cases of causal overdetermination. The *legal liability-responsibility* strings can help to clarify that in most cases (though not all), a negative outcome has to actually have occurred before the question of who is liable is raised. Moreover, a case actually has to be brought.[29] Even so, with further refinement, the liability strings can also support thinking through which actors could or would be liable for adverse consequences to inform forwards-looking best practice. The *moral responsibility* strings can help to support conversations about people being at risk of being unjustly held responsible for incidents and accidents (Lawton et al., 2023; Elish, 2019) in virtue of their contributions being non-voluntary (e.g., individual engineers acting under extreme pressure, or operators having no meaningful capacity to

[29] In 2019, prosecutors said that Uber ATG was not criminally liable in the crash. Uber ATG also settled the civil case brought by the victim's family out of court. Uber preferred to use the word "resolved", in place of settled, perhaps for public relations reasons since they did not admit liability, (Fulbrook, 2021), although settlements which compromise a dispute often contain provisions that a party does not accept liability. Amongst other things, this settlement represents a missed opportunity to test and develop important questions of civil liability for incidents involving AI-enabled systems. To note, the safety driver was charged with negligent homicide in 2020, and this year pleaded guilty to endangerment, a reduced charge; she has been sentenced to three years of supervised probation (Shepardson, 2023).

intervene before harm is caused). These strings may also help to support reasoning about 'responsibility gaps', by providing a structure to trace back to decision and action-points when actors voluntarily committed to the risk of delegating to an AI-enabled system, despite limited control over its decisions* and actions* during operation.

There are two next steps for the work.

First, we intend to lay out and present the responsibility strings in a visual map or graph which could be used to pick out different responsibility "pathways" between actors and occurrences. Such a map would function as a heuristic for allocating prospective role-responsibilities to actors, as well as for tracing legal liability and moral responsibility retrospectively – or for thinking in advance to whom these would be traced – after an incident or accident has occurred. Though this paper stops at 'unravelling responsibility' as a precursor to that connective work, there are some lines of enquiry and principles that can guide progress here.

First, there is the question of the causal relations and dependencies between different kinds of *O*. For example, actions cause consequences, and consequences can often cause us to go and make decisions. Deeper reflection on causal chains will be central to the mapping exercise. Second, there are standard internal connections between types or senses of responsibility. For example, causal responsibility is connected to the other three kinds of responsibility. When one has a role-responsibility, one is generally responsible for producing (i.e., causing) a good outcome or for preventing a bad one (Cane, 2002, p. 31). Where legal liability requires harm or damage to have occurred, causation is also almost universally required for legal liability.[30] And though causal responsibility should not be identified with moral responsibility, it is a necessary condition for moral responsibility. Other general principles which will help to inform the map concern the conditions for the kind of responsibility in question. For example, on the account given, the condition for *A* to bear moral responsibility as attributability for *O* is that *A* voluntarily caused *O*, where exculpations would include acting under extreme pressure, duress or coercion as well as acting in ignorance.

These three sets of consideration (causal connections between occurrences, connections between the kinds of responsibility, and conditions and exceptions for each kind of responsibility) are the basis upon which we will start to consider how the strings may combine and come apart when designing a responsibility 'map' for charting responsibility journeys for occurrences in which AI is involved.

Second, we intend to use the vocabulary set out in this paper to inform and structure responsibility assurance cases. Assurance cases are structured arguments which provide defeasible reasons for believing that a claim or goal regarding a system property of interest, typically safety, is true (SCSC-141C, 2021; Goodenough, Weinstock & Klein, 2012). The assurance case methodology could be adapted to the question of responsibility, for instilling confidence in stakeholders, for example, that responsibility (in one or more senses) for *system outputs* can be located and that responsibility for *human or institutional decisions and actions*, and their *consequences*, can be traced. The breakdown provided in this paper furnishes a starting point for such undertakings. As an indication of

[30] Rare exceptions include situations of factual uncertainty and exceptions which have evolved in the context of toxic torts, particularly mesothelioma (e.g. *Fairchild v Glenhaven Funeral Services Ltd* [2002] UKHL 22; [2003] 1 AC 32).

how this is starting to be pursued, see (Ryan Conmy et al., 2023b), which introduces a responsibility assurance argument with a top goal concerning role-responsibility.

More work needs to be done, but this paper, in unravelling and distilling the rich concept of responsibility into a structured set of meanings of 'A is responsible for O', represents a start. It is an attempt at clearing the knots and weeds and tangles in the language we use to talk about responsibility for AI.

–

REFERENCES

Aristotle (2002) *Nicomachean ethics.* Oxford University Press.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989.

Askell, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development. arXiv preprint arXiv:1907.04534.

SCSC-141C. (2021). The Assurance Case Working Group: SCSC-141C. Goal Structuring Notation Community Standard (Version 3) https://scsc.uk/r141C:1?t=1

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, 279, 103201.

Cane, P. (2002). *Responsibility in law and morality*. Bloomsbury Publishing.

Chesterman, S. (2020). Artificial intelligence and the limits of legal personality. *International & Comparative Law Quarterly*, 69(4), 819-844.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 26(4), 2051-2068.

Cooper, A. F., Moss, E., Laufer, B., & Nissenbaum, H. (2022, June). Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 864-876).

Dekker, S. W., & Breakey, H. (2016). 'Just culture:'Improving safety by achieving substantive, procedural and restorative justice. *Safety Science*, 85, 187-193.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way* (Vol. 2156). Cham: Springer.

Elish, M.C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5, 40-60.

Fairchild v Glenhaven Funeral Services Ltd [2002] UKHL 22; [2003] 1 AC 32

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

Feinberg, J. (1970) *Doing & deserving; essays in the theory of responsibility*. Princeton University Press.

Fourth Motor Insurance Directive, Directive 2000/26/EC of the European Parliament and of the Council of 16 May 2000 on the approximation of the laws of the Member States relating to insurance against civil liability in respect of the use of motor vehicles and amending Council Directives 73/239/EEC and 88/357/EEC.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14, 349-379

French, P. (1984). *Collective and corporate responsibility*. Columbia University Press.

French, P. (ed). (1998). *Individual and collective responsibility*. Schenkman Books.

Giliker, P. (2018). Comparative Law and Legal Culture: Placing Vicarious Liability in Comparative Perspective. *The Chinese Journal of Comparative Law,* 6, 265-293.

Goodenough, J., Weinstock, C., Klein, A.(2012) Toward a theory of assurance case confidence. Carnegie Mellon University Report https://apps.dtic.mil/sti/pdfs/ADA609836.pdf

Hart, H.L.A., (1968/2008) *Punishment and responsibility: Essays in the philosophy of law* (2nd edn). Oxford University Press.

Horder, J. (2022). *Ashworth's Principles of Criminal Law* (10th edn). Oxford University Press.

Johnson, D. (2006) Computer systems: moral entities but not moral agents. *Ethics and Information Technology*, 8, 195–204.

Jones, C., Thornton, J., & Wyatt, J. C. (2023). Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. *Medical Law Review* https://doi.org/10.1093/medlaw/fwad013

Kutz, C. (2004). Responsibility, in Coleman, Himma & Shapiro (eds), *The Oxford Handbook of Jurisprudence and Philosophy of Law* (2004; online edn, Oxford Academic, 18 Sept. 2012), https://doi.org/10.1093/oxfordhb/9780199270972.013.0014

Lawton, T., Morgan, P., Porter, Z., Hickey, S., Cunningham, A., Hughes, N., Iacovides, J., Jia, Y., Sharma, V., Habli, I. (2023). Clinicians risk becoming 'liability sinks' for Artificial Intelligence. *Authorea*. https://d197for5662m48.cloudfront.net/documents/publicationstatus/137543/preprint_pdf/57fb8768141d5decde63d7ffa068c093.pdf

List, C. & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175-183

Morgan, P. (ed.) (2023). *Tort Liability and Autonomous System Accidents: Common Law and Civil Law Perspectives*. Edward Elgar.

Morgan, P. (2024). Tort Law and AI – Vicarious Liability, in Lim and Morgan (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence*. Cambridge University Press. (forthcoming).

National Transportation Safety Board (2019) Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018, NTSB/HAR-19/03. https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2(1), 25-42.

NIST/CNSSI 4009-2015 https://csrc.nist.gov/glossary/term/commercial_off_the_shelf#:~:text=Definitions%3A,license%20to%20the%20general%20public.

Noordhof, P. (2020). *A variety of causes*. Oxford University Press.

Ormerod, D.C. & Perry, D. (2021). *Blackstone's Criminal Practice 2021*. Oxford University Press.

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., Guston, D. (2013). A framework for responsible innovation, in Owen, Bessant & Heinz (eds.) *Responsible innovation: managing the responsible emergence of science and innovation in society*, 27-50. Wiley.

Ozturk, B., Lawton, T., Smith, S., & Habli, I. (2023). Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning. In *Caring is Sharing–Exploiting the Value in Data for Health and Innovation* (38-42). IOS Press.

Pagallo, U. (2013). *The Laws of Robots: crimes, contracts, and torts.* Springer.

Pesch, U. (2008). Administrators and accountability: The plurality of value systems in the public domain. *Public Integrity*, 10(4), 335-344.

Peters, D., Vold, K., Robinson, D. and Calvo, R.A. (2020). Responsible AI—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, 1(1), 34-47.

Pettit, P. (2007). Responsibility incorporated. *Ethics*, 117, 171-201.

Porter, Z., Zimmermann, A., Morgan, P., McDermid, J., Lawton, T., & Habli, I. (2022). Distinguishing two features of accountability for AI technologies. *Nature Machine Intelligence*, 4(9), 734-736.

Product Liability Directive, Council Directive 85/374/EEC of 25 July 1985 On the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products

Rubel, A., Castro, C. & Pham, A. (2019). Agency laundering and information technologies. *Ethical Theory and Moral Practice*, 22(4), 1017-1041.

Ryan Conmy, P., Ozturk, B., Lawton, T., Habli, I. (2023a). The Impact of Training Data Shortfalls on Safety of AI-based Clinical Decision Support Systems. SAFECOMP 2023 (*forthcoming*)

Ryan Conmy, P. M., Porter, Z., Habli, I., McDermid, J. A. (2023b). Safety engineering, role responsibility and lessons from the Uber ATG Tempe Accident. In *First International Symposium on Trustworthy Autonomous Systems* (TAS'23). Association for Computing Machinery.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., Lennon, M. (2020). Principles to practices for responsible AI: closing the gap. arXiv preprint arXiv:2006.04707

Shepardson, D. (2023). Backup driver in 2018 Uber self-driving crash pleads guilty. Reuters, Friday 28 July 2023 Backup driver in 2018 Uber self-driving crash pleads guilty (msn.com)

Shoemaker, D. (2013). On criminal and moral responsibility. *Oxford Studies in Normative Ethics,* 3, 154-178.

Shoemaker, D. (2015). *Responsibility from the margins*. Oxford University Press.

Smiley, M. (2022). Collective responsibility. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2023/entries/collective-responsibility

Strawson, P. F. (1962). Freedom and Resentment, in Proceedings of the British Academy, 48: 1–25. Reprinted Fischer and Ravizza (eds.), *Perspectives on Moral Responsibility* 1993, 45–66.

Talbert, M. (2022). Moral responsibility. *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/fall2022/entries/moral-responsibility

Thoma, J. (2022). Risk imposition by artificial agents: The moral proxy problem, in Vöneky, Kellmeyer, Müller & Burgard (eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*. Cambridge University Press.

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review*, 74(4), 905-916.

Tobia, K., Nielsen, A., & Stremitzer, A. (2021). When does physician use of AI increase liability? *Journal of Nuclear Medicine*, 62(1), 17-21.

Tuomela, R. (1989). Actions by collectives. *Philosophical Perspectives*, 3, 471-496.

Turner, J. (2019). *Robot Rules: Regulating Artificial Intelligence.* Palgrave.

Vallor, S. (2023). Edinburgh Declaration on responsibility for Responsible AI, *Medium,* July 14 2023. https://medium.com/@svallor_10030/edinburgh-declaration-on-responsibility-for-responsible-ai-1a98ed2e328b

Van de Poel, I., Royakkers, L.M., Zwart, S.D., De Lima, T. (2015). *Moral responsibility and the problem of many hands*. Routledge.

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & SOCIETY*, 36(2), 487-497.

Verdicchio, M., & Perin, A. (2022). When doctors and AI interact: on human responsibility for artificial risks. *Philosophy & Technology*, 35(1), 11.

Vincent, N. A. (2011). A structured taxonomy of responsibility concepts. In Vincent, van de Poel & van den Hoven (eds) *Moral responsibility: Beyond free will and determinism*, 15-35. Springer Netherlands.

Von Schomberg, R. (2013). A vision of responsible research and innovation, in Owen, Bessant & Heinz (eds.) *Responsible innovation: managing the responsible emergence of science and innovation in society*, 51-74. Wiley.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press

Walton, C.T., Armitage, M., Cooper, R., Hyde, R., Kramer, P., Todd., S. (eds.) (2018) *Charlesworth & Percy on Negligence 14th Ed*. Sweet and Maxwell.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, 24(2), 227-248.

Zwart, S. D., Van de Poel, I., Van Mil, H., Brumsen, M. (2006). A network approach for distributing ethical issues in research and development. *Science and Engineering Ethics*, 12(4), 663-684.

APPENDIX

	A	**Is role-responsible for**	***O***
R1.	AI-enabled system	Is (task)role-responsible for	Decision*
R2.	AI-enabled system	Is (task)role-responsible for	Action*
R3.	AI-enabled system	Is (task)role-responsible for	Omission*
R4.	AI-enabled system	Is (task)role-responsible for	Decision[31]
R5.	AI-enabled system	Is (task)role-responsible for	Action
R6.	AI-enabled system	Is (task)role-responsible for	Omission
R7.	AI-enabled system	Is (task)role-responsible for	Consequence
R8.	Individual Human	Is (task)role-responsible for	Decision*
R9.	Individual Human	Is (task)role-responsible for	Action*
R10.	Individual Human	Is (task)role-responsible for	Omission*
R11.	Individual Human	Is (task)role-responsible for	Decision
R12.	Individual Human	Is (task)role-responsible for	Action
R13.	Individual Human	Is (task)role-responsible for	Omission
R14.	Individual Human	Is (task)role-responsible for	Consequence
R15.	Individual Human	Is (moral obligation)role-responsible for	Decision*
R16.	Individual Human	Is (moral obligation)role-responsible for	Action*
R17.	Individual Human	Is (moral obligation)role-responsible for	Omission*
R18.	Individual Human	Is (moral obligation)role-responsible for	Decision
R19.	Individual Human	Is (moral obligation)role-responsible for	Action
R20.	Individual Human	Is (moral obligation)role-responsible for	Omission
R21.	Individual Human	Is (moral obligation)role-responsible for	Consequence
R22.	Individual Human	Is (legal duty)role-responsible for	Decision*
R23.	Individual Human	Is (legal duty)role-responsible for	Action*
R24	Individual Human	Is (legal duty)role-responsible for	Omission*
R25.	Individual Human	Is (legal duty)role-responsible for	Decision

[31] The idea in R4-R6 is that it might be part of the AI-enabled system's task to get a human to make a decision or perform an action (e.g. when issuing a transition demand, an AV is (task)role-responsible for getting the human driver to take on the action of driving or steering the vehicle).

R26.	Individual Human	Is (legal duty)role-responsible for	Action
R27.	Individual Human	Is (legal duty)role-responsible for	Omission
R28.	Individual Human	Is (legal duty)role-responsible for	Consequence
R29.	Institution	Is (task)role-responsible for	Decision*
R30.	Institution	Is (task)role-responsible for	Action*
R31.	Institution	Is (task)role-responsible for	Omission*
R32.	Institution	Is (task)role-responsible for	Decision
R33.	Institution	Is (task)role-responsible for	Action
R34.	Institution	Is (task)role-responsible for	Omission
R35.	Institution	Is (task)role-responsible for	Consequence
R36.	Institution	Is (legal duty)role-responsible for	Decision*
R37.	Institution	Is (legal duty)role-responsible for	Action*
R38.	Institution	Is (legal duty)role-responsible for	Omission*
R39.	Institution	Is (legal duty)role-responsible for	Decision
R40.	Institution	Is (legal duty)role-responsible for	Action
R41.	Institution	Is (legal duty)role-responsible for	Omission
R42.	Institution	Is (legal duty)role-responsible for	Consequence

Table 3: Strings of '*A* is role-responsible for *O*'

	A	**Is causally responsible for**	***O***
C1.	AI-enabled system	Is causally responsible for	Decision*
C2.	AI-enabled system	Is causally responsible for	Action*
C3.	AI-enabled system	Is causally responsible for	Omission*
C4.	AI-enabled system	Is causally responsible for	Decision
C5.	AI-enabled system	Is causally responsible for	Action
C6.	AI-enabled system	Is causally responsible for	Omission
C7.	AI-enabled system	Is causally responsible for	Consequence
C8.	Individual Human	Is causally responsible for	Decision*
C9.	Individual Human	Is causally responsible for	Action*

C10.	Individual Human	Is causally responsible for	Omission*
C11.	Individual Human	Is causally responsible for	Decision
C12.	Individual Human	Is causally responsible for	Action
C13.	Individual Human	Is causally responsible for	Omission
C14.	Individual Human	Is causally responsible for	Consequence
C15.	Institution	Is causally responsible for	Decision*
C16.	Institution	Is causally responsible for	Action*
C17.	Institution	Is causally responsible for	Omission*
C18.	Institution	Is causally responsible for	Decision
C19.	Institution	Is causally responsible for	Action
C20.	Institution	Is causally responsible for	Omission
C21.	Institution	Is causally responsible for	Consequence

Table 4: Strings of '*A* is causally responsible for *O*'

	A	**Is legal liability-responsible for**	***O***
L1.	Individual Human	Is (criminal)liability-responsible for	Consequence
L2.	Individual Human	Is (civil)liability-responsible for	Consequence
L3.	Institution	Is (criminal)liability-responsible for	Consequence
L4.	Institution	Is (civil)liability-responsible for	Consequence

Table 5: Strings of 'A is legal liability-responsible for O'

	A	**Is morally responsible for**	***O***
M1.	Individual Human	Is (attributability)morally responsible for	Decision*
M2.	Individual Human	Is (accountability)morally responsible for	Decision*
M3.	Individual Human	Is (attributability)morally responsible for	Action*
M4.	Individual Human	Is (accountability)morally responsible for	Action*
M5.	Individual Human	Is (attributability)morally responsible for	Omission*
M6.	Individual Human	Is (accountability)morally responsible for	Omission*
M7.	Individual Human	Is (attributability)morally responsible for	Decision
M8.	Individual Human	Is (accountability)morally responsible for	Decision

M9.	Individual Human	Is (attributability)morally responsible for	Action
M10.	Individual Human	Is (accountability)morally responsible for	Action
M11.	Individual Human	Is (attributability)morally responsible for	Omission
M12.	Individual Human	Is (accountability)morally responsible for	Omission
M13.	Individual Human	Is (attributability)morally responsible for	Consequence
M14.	Individual Human	Is (accountability)morally responsible for	Consequence

Table 6: Strings of '*A* is morally responsible for *O*'