Unravelling Responsibility for AI

ZOE PORTER, Department of Computer Science, University of York, UK
PHILIPPA RYAN, Department of Computer Science, University of York, UK
PHILLIP MORGAN, York Law School, University of York, UK
JOANNA AL-QADDOUMI, York Law School, University of York, UK
BERNARD TWOMEY, Department of Computer Science, University of York, UK
JOHN MCDERMID, Department of Computer Science, University of York, UK
IBRAHIM HABLI, Department of Computer Science, University of York, UK

It is widely acknowledged that we need to establish where responsibility lies for the outputs and impacts of AI-enabled systems. But without a clear and precise understanding of what 'responsibility' means, deliberations about where responsibility lies will be, at best, unfocused and incomplete and, at worst, misguided. To address this concern, this paper draws upon central distinctions in philosophy and law to clarify the concept of responsibility for AI for policymakers, practitioners, researchers and students from non-philosophical and non-legal backgrounds. Taking the three-part formulation 'Actor A is responsible for Occurrence O,' the paper unravels the concept of responsibility to clarify that there are different possibilities of *who* is responsible for AI, *senses* in which they are responsible, and *aspects of events* they are responsible for. Criteria and conditions for fitting attributions of responsibility in the core senses – causal responsibility, role-responsibility, liability responsibility and moral responsibility – are articulated to promote an understanding of when responsibility attributions would be inappropriate or unjust. The analysis is presented with a graphical notation to facilitate informal diagrammatic reasoning and discussion about specific cases. It is illustrated by application to a scenario of a fatal collision between an autonomous AI-enabled ship and a traditional, crewed vessel at sea.

1 INTRODUCTION

The importance of establishing where responsibility lies for the outputs and impacts of AI-based systems has been emphasised in the majority of ethical principles proposed for AI [39, 58], by the OECD [76], and in UNESCO's global standard on the ethics of AI [106]. It is also recognised, in heterogeneous ways, in emerging AI regulation and legislation worldwide [16, 17, 57, 107]. Whether forwards-looking responsibility in the form of roles and duties to ensure that AI-powered systems do not cause harm or violations of fundamental rights, or backwards-looking liability and moral accountability for AI-occasioned harms or wrongs after they have occurred [11, 32, 50, 110], policymakers and stakeholders in AI governance need to ensure that powerful actors do not obscure or evade their responsibility for the impacts of these technologies [89]. Meanwhile, clarity on the location of responsibility for AI, and specifically for systems enabled by machine learning (ML), will likely instil public trust in the use of these technologies and encourage diligence and cooperation amongst practitioners and stakeholders.

The project of pinpointing responsibility for AI is, however, inherently difficult. AI presents a problem of 'many hands', whereby the sheer number of actors influencing a system's behaviour and its consequences makes attributing responsibility to discernible individuals a highly complex undertaking [9, 14, 20, 73, 103, 111]. And when AI is deployed

Authors' addresses: Zoe Porter, zoe.porter@york.ac.uk, Department of Computer Science, University of York, UK; Philippa Ryan, philippa.ryan@york.ac.uk, Department of Computer Science, University of York, UK; Phillip Morgan, York Law School, University of York, UK; Joanna Al-Qaddoumi, York Law School, University of York, UK; Bernard Twomey, Department of Computer Science, University of York, UK; John McDermid, Department of Computer Science, University of York, UK; Ibrahim Habli, Department of Computer Science, University of York, UK.

in increasingly autonomous systems, which can directly harm people independently of human input, responsibility gaps and liability gaps arguably arise [46, 47, 55, 61, 69, 72, 84, 87, 105]. One practical answer to the problem of responsibility gaps in this context is to limit the autonomy of the systems by ensuring that there is always a human 'in-the-loop', but this, in turn, can mean that the nearest human operator disproportionately and unfairly absorbs blame and liability for overall system malfunctions over which they have limited control or knowledge [34, 64]. These challenges have prompted philosophical and legal replies [5, 45, 56, 61, 72, 75].

Underlying this normative complexity there is the further difficulty that responsibility is a rich concept, with distinct but overlapping meanings which are prone to confusion. Failure to be clear and specific about what responsibility *means* can lead to incomplete and misguided reasoning about where responsibility *lies*. Unravelling the concept of responsibility is therefore purpose of this paper. Our objective is to provide a sound conceptual foundation for policymakers, practitioners, researchers and students, and specifically those who are not responsibility specialists, to reason about responsibility for AI for themselves, whilst avoiding some of the conceptual tangles that can arise in debates and discussions of responsibility for AI.

At the heart of the work is the division of responsibility into four senses – four different *ways of being responsible*: causal, role, liability, and moral responsibility. This derives from the legal philosopher H.L.A Hart's [52] taxonomy of the senses of responsibility, and has already been used in some general, conceptual introductions to responsibility for AI [50, 120]. This paper extends that conceptual clarification by setting the four core senses of responsibility within the relational context of *who* (or *what*) may be responsible in these senses and *for what* part of the event involving an AI-based system. We also present criteria and conditions for being rightly responsible in these senses to engender greater understanding of when responsibility attributions may be inappropriate or unjust. And we offer an informal graphical notation of the analysis and a heuristic method for modelling the distribution of responsibility for AI in specific cases to support discussion.

The paper is structured as follows. In Section 2, the concept of 'responsibility' is systematically decomposed. Each of the three parts of the statement 'Actor A is responsible for Occurrence O' are broken down into subcategories (the three parts being: A; is responsible for; and O). Criteria and conditions for the types of ways of being responsible for – causal, role, legal and moral responsibility – are articulated. The analysis is represented in a graphical notation, drawing on Ryan et al. [90]. In Section 3, we apply the conceptual analysis and notation to a maritime responsibility scenario in which multiple causes, including sensors degraded by salt, have contributed to a fatal collision between an autonomous and traditional vessel. In Section 4, uses of the work are discussed. There are three main intended audiences. First, policymakers who could gain from the basis it provides for extending questions of responsibility for AI beyond 'who?' to 'what for?' and 'how?'. Second, practitioners and technical researchers, particularly AI safety engineers, who could find that laying out how different actors bear different responsibilities for different occurrences adds a useful dimension to foundational research or risk management. Third, students from diverse disciplines exploring the moral and legal implications of AI technologies, and their teachers, who could find this a useful reaching material since it summarises and distils a body of literature and can be applied to specific cases.

2 UNRAVELLING RESPONSIBILITY: ACTOR A IS RESPONSIBLE FOR OCCURRENCE O

'What does 'responsibility' mean?'. This apparently simple question is surprisingly difficult to answer. The etymology is that it comes from the Latin *respondere*, meaning 'to answer' [30, 67]. This is rather broad, and is effectively just synonymy. Dictionary definitions, on the other hand, tend be narrower. The Oxford English Dictionary online, for

example, defines 'responsibility' as: *"The state or fact of being in charge of or of having a duty towards a person or thing; obligation"*[79]. This only covers what we call 'role-responsibility'.

Responsibility is an area of inquiry in philosophy and law, and general statements from these fields can furnish a more precise answer to the question. 'Responsibility' describes a kind of relation between an actor and an occurrence [98], whether that is an action, an omission or an outcome. Responsibility is ascribed *to* actors. Actors are responsible *for* occurrences. Building on this, in Section 2 we systematically break down the three parts of the three-part statement, 'Actor A is responsible for Occurrence O' to clarify the different kinds or senses of responsibility that can be attributed to different kinds of actor for different kinds of occurrences involving AI. We omit an additional relational element of responsibility, namely, the responsibility of actors *to* others [15, 109], partly to manage initial complexity, and partly because it is often implicit in responsibility attributions (a duty of care, for example, is a duty towards one's 'neighbours', and moral accountability is accountability to the moral community).

2.1 Actor A

In the majority of cases and incidents involving AI, multiple actors will be involved. We delineate three kinds of Actor A: an AI-based system; an individual human; an institution. Not all kinds of actor can be responsible in the same way. More specifically, we proceed on the basis that AI-based systems cannot be legally and morally responsible, as discussed below. The three subcategories of Actor A are shown in Figure 1.

Fig. 1. The three subcategories of Actor A

To give some examples, an *AI-based system* could be an ML-based radiology imaging system, an or an autonomous vehicle. Relevant *individual humans* would include data scientists, designers, programmers, safety engineers, users, individual operators, public officials, and corporate managers. *Institutions* would include the software development companies, manufacturers, public service operators, and national or international regulators and certification bodies.

One might one ask how *human-AI teams* should be categorised on this tripartite scheme. For present purposes, we suggest that these could be dealt with by combining an AI-enabled system and a human individual and thinking about their respective responsibility relations to a particular occurrence. One might also ask how to account for *swarms of AI-enabled systems*. These could be categorised either as an aggregate of individual systems, or the swarm as a whole could be regarded as a single AI-enabled system. One might also point to *different kinds of relations between individuals* both within and outside of institutions. For example, within institutions, individuals may act together or a few actors may be delegated actions to perform on behalf of the whole group [21, 68]; both options are covered under the single subcategory of 'institution' but they are not delineated. To be clear, the work is intended to set out the basic conceptual framework that enables more fine-grained discussions such as these, and their practical, moral and legal implications, but not to prescribe too many conclusions.

It is important to be clear that AI systems are not legal persons. As such, they cannot have legal duties or be held liable for harms or injury that they cause. The standard position in legal scholarship is to advocate against the legal personhood of current and near-term AI systems [13, 72]. By contrast, most natural persons are legally responsible for what they do. Since most corporations and other institutions are legal persons, these are too. On our analysis, individual humans and institutions can therefore have have legal duties and be held liable.

This paper takes the position that AI-powered systems are not moral agents and therefore cannot have moral duties or be morally responsible for what they do. This is the standard view in the current literature. Various reasons have been given for it, coalescing on the argument that the machines lack a necessary property for moral agency such as sentience [112, 113], or consciousness [94], or understanding [15], or the capacity to act from their own reasons, motives or intentions [59, 88, 113]. Even amongst those in the literature who have tried to make conceptual space for the notion of an artificial moral agent, either by arguing for non-anthropomorphic moral agency [40, 48, 115] or by arguing that some machines actually possess relevant features they have been considered to lack, such as intentionality [86], most, but not all [65], argue that artificial agents, or AI-based systems, cannot go so far as to be morally responsible.

Philosophers debate whether or not institutions, such as corporations and governments, are moral agents [95]. We adopt the standard position that institutions are moral agents. Consistency requires that the reasons for ruling AI-based systems out of the class of moral agents should not be relaxed when admitting institutions into that class [65]. Taking it broadly to be the case, based on an Aristotelian tradition, that a moral agent can act voluntarily and foresee the consequences [22], it is not clear that institutions - as supra-entities which transcend their individual human members - should have these psychological capacities and be ruled in while AI-based systems are ruled out, even though this has been argued by many in business ethics and moral philosophy [6, 44, 66, 82]. Let us say, then, that institutional moral agency is to be understood in the weaker sense that it ultimately reduces to the moral agency of its individual members, who do have capacities to choose and understand. Institutional agency is the aggregate of this individual agency in a specific context [10, 68]. On this weaker sense of the moral agency of institutions, institutions can have moral duties and be morally responsible for their actions, omissions and their consequences.

To summarise, on our decomposition in Section 2.2.2 and 2.2.3 below, AI-based systems cannot have legal duties or be held liable, while individual humans and institutions can both have legal duties and be held liable. And in Section 2.2.2 and 2.2.4, they cannot have moral duties or be morally responsible, but individual humans and institutions can have both moral duties and be morally responsible. One implication of the second point is that talk of transferring moral duties or moral responsibility to AI-based technologies is illegitimate. Indeed, such talk could be used by individual human and institutional actors to avoid their own moral duties or diminish their own moral responsibility for what occurs, shifting it onto AI-based systems [101].

2.2 Is responsible for

'Responsibility' is a term with a plurality of senses. In common with some general, conceptual introductions to responsibility for AI [50, 120], we adapt the legal philosopher H.L.A Hart's [52] taxonomy of the senses of responsibility. Our adaptation of Hart's analysis is shown in Table 1.

Distinguishing these senses enables us to be specific and focused in discussions about where responsibility lies for AI. Sometimes these senses, or the relations between them, are confused unwittingly. It is common, for example, for people to over-identify causal responsibility with moral responsibility, as we discuss below in Section 2.2.1. Conceptual confusion and ambiguity can also be deliberately traded on: it enables actors to equivocate, particularly to avoid liability or moral accountability. Specificity can also help us to articulate where actors are in danger of being scapegoated. For

Table 1. The main senses of 'is responsible for'

Sense of responsibility	**Description**
Causal responsibility	A is a cause of O
Role-responsibility	A has tasks, moral duties or legal duties that attach their role
Legal liability-responsibility	A is liable to legal sanction for O
Moral responsibility	A is an author of O (moral responsibility as attributability)
	A is liable to moral sanction for O (moral responsibility as accountability)

example, role-responsibility does not entail liability or moral accountability, even though it is a necessary condition of it on our analysis; as such, the legal or moral sanction of role-holders who did not meet the other conditions would be a form of scapegoating. This is discussed further in in Section 2.2.2.

Those familiar with Hart's taxonomy will note some differences. Causal responsibility has been placed first rather than second in the list. This is not to attach any particular normative significance to causal responsibility. Another difference is that moral responsibility is expanded to include a distinction from Watson [116], on the difference between moral responsibility as attributability and accountability. Finally, Hart's taxonomy also identifies 'capacity-responsibility' which we omit. 'Capacity-responsibility' is the possession of psychological capacities such as understanding, reasoning and self-control. It is a basic criterion for a person's being morally responsible and, in most cases, liable for what they do. In this paper, the capacity-responsibility of AI actors who are natural persons is assumed.

2.2.1 Is causally responsible for. Within this paper, we take causal responsibility to be another way of referring to causality. For maximum generality, this paper proposes that, to be causally responsible for O, an actor needs only to have been *a* cause of O, not *the only* or *the most important cause* of O. This is an inclusive notion of causality[74], and it precludes the necessity of taking a position in debates about what makes something *the most important* cause of an occurrence [24, 49, 60, 63, 102, 118]. Many conditions of an inclusive notion of causation - conditions for being *a* cause - have been proposed [4]. We suggest using the NESS condition. The NESS condition is a cause must be a Necessary Element of a Sufficient Set [54, 119]. As Wright puts it, this is a "test for causal contribution that is applicable to the entire spectrum of causation cases." [119]. This serves our purposes well. In the maritime case study in 3, several actors' omissions are necessary elements of a sufficient set of conditions for the fatal collision. The NESS condition of what constitutes a cause is a threshold condition: if the actor's contribution to the occurrence meet the NESS test, causation holds, and therefore, at least on this paper's analysis, the actor bears a relation of causal responsibility to the occurrence.

Further, it is not just actors that can be causally responsible [52]. Other occurrences *O*, facts and events can also cause occurrences. These relations can be depicted by showing causal responsibility arrows from one occurrence to another, as in Figure 2. The graphical notation for an A's causal responsibility for an O is shown in Figure 2.

Causal responsibility is foundational to the other senses of responsibility. It is connected to role-responsibility. As will be seen in Section 2.2.2 below, role-responsibility describes what people have a duty or obligation to bring about or to prevent, both of which are causal notions. Causal responsibility is a necessary (but not a sufficient) condition for liability and moral responsibility. Without causal responsibility for occurrences, actors are not even in the ballpark for liability and moral responsibility, which are the senses we are most often concerned about: Who is liable? Who should pay compensation? Who should we blame? But we need to be very clear what 'in the ballpark' means here, since this is prone to misunderstanding, with negative normative consequences. It means that A cannot be liable or morally responsible *without* a causal connection to the occurrence (although in cases of vicarious liability, this is a

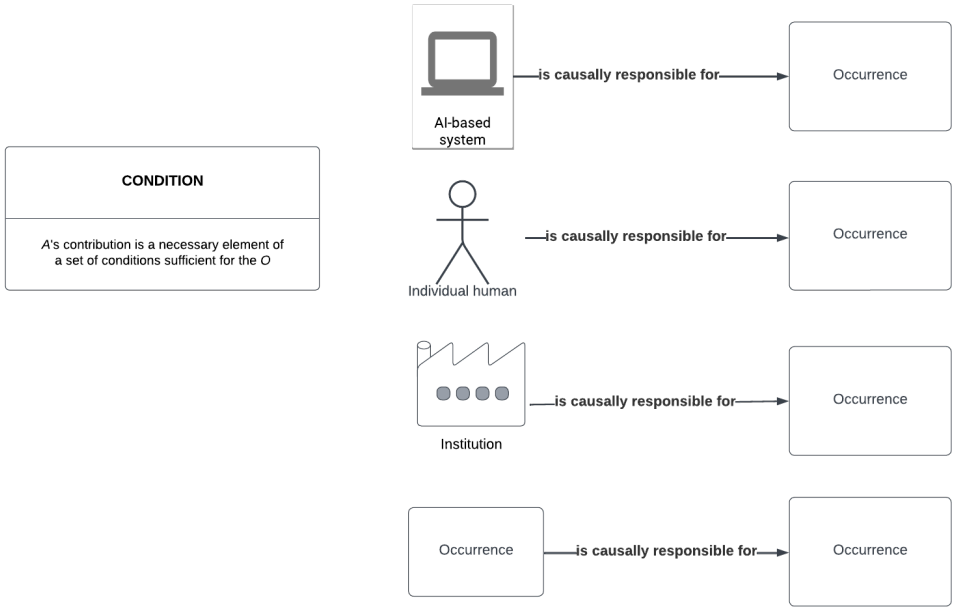

Fig. 2. Actor A (or Occurrence O) is causally responsible for Occurrence O

little more complicated), but having this causal connection does not *entail* that A is liable or morally responsible. Other conditions must also be met. These conditions are set out in Sections 2.2.3 and 2.2.4.

Understanding this can help to sharpen people's awareness that pointing to the actor with the nearest causal connection to a harm, wrong, damage or injury from an AI-based system – typically the human in-the-loop – and arguing that they, and they alone, should therefore be 'on the hook' for it in a moral or legal sense can be a way of scapegoating that individual, whilst diminishing the liability to legal or moral sanction which others should bear. This is a documented concern in the automotive sector [34], in defence [28] and in healthcare [64, 99], where human-AI teaming models mean that the the human in-the-loop is typically required to intervene to prevent an AI-enabled system from causing an accident or harm.

In the world of safety science, the relation of causation to other ways of being responsible is currently at risk of being confused, to the extent that 'cause' is seen almost as a 'trigger word' that should be avoided at all costs in an effective safety culture [26]. In a recent NHS England document on patient safety incident investigations, for example, investigators are advised not to use the word 'cause' at all, because it is *"stongly associated with blame and liability"* [36]. But we cannot dispense with causal talk in incident investigation otherwise we will not understand how harms and injuries came about; we simply need to be clear that causation, or causal responsibility, does not entail blame or liability.

2.2.2 Is role-responsible for. Another way in which an actor might be responsible is by having a tasks, duties and functions attached to their specific role in an organisation or society. Responsibility in this sense is connected to policies of Responsible Research and Innovation [80] and Responsible AI [1, 31, 78, 93], which concern actors taking seriously their duties to design, engineer, manufacture, use and govern innovative technologies in ways that steer them towards societally desirable goals and which do not, amongst other things, compound existing inequalities [33, 78, 114].

For specificity, we identify three kinds of role-responsibility an actor might bear: task responsibility; moral duties; and legal duties. These are shown in Figure 3 below.

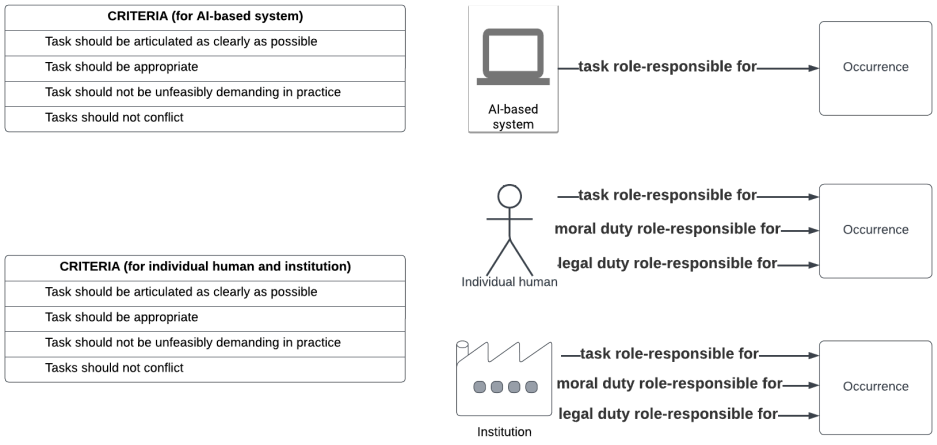

Fig. 3. Actor A is role-responsible for Occurrence O

Because AI-based systems are neither moral agents, on our account, nor legal persons, they cannot have either moral or legal duties, only tasks that they carry out. *Task responsibilities* are those functions and duties that attached to an actor's role. Typically, these task responsibilities are to perform certain actions which are intended to bring about or prevent an occurrence. These are inherently causal notions. One salient question in the context of responsibility for AI is the question of whose role-responsibility it is to articulate the AI-based system's tasks. Typically, this is the engineer but, given the context-dependent nature of the systems and their uses, it would ideally include the perspective of operators, end users and regulators. *Moral duties* are what morality requires of us. Broadly speaking, we have a moral duty not to harm others unjustly (because people have a right not to be harmed or wrong) and, in some circumstances, a moral duty to come to the aid of others [41]. In reality, the contours and limits of our moral duties are often contested, and are not always clear. *Legal duties* are set out in law, such as the EU's AI Act (which is not yet in force) with its duty on AI producers of high-risk AI systems to carry out pre-deployment conformity assessments. A legal duty may be a *duty of care* not to cause actionable harm by failing to meet the standard of a reasonable person carrying out the same function [81], or an *absolute duty* that must be adhered to whatever the effort, time or cost, or another legal standard.

We suggest the following criteria for the effective discharge of a role-responsibility - the idea being that, if any of these criteria are not met, then it would be inappropriate to impose the role-responsibility in question upon the actor and, in the case of individual humans and institutions as moral agents and legal persons, it would be unfair to do so. First, the role-responsibility in question should be defined or articulated as clearly as possible. Ambiguous role-responsibilities can facilitate scapegoating individuals, because they may be held accountable for failing to do something it was unclear that they should do; by the same token, ambiguity and vagueness can facilitate responsibility-evasion, because actors may deliberately equivocate about what it was their task or duty to bring about or prevent. Second, the role-responsibility should be appropriate to the context. In other words, it should discharge a particular need – the designer of a system intended solely to tolerate extremely cold conditions, for example, should not be tasked to design a system that can

tolerate extreme heat – and the task should be proportionate to that need. Third, the role-responsibility should not be unfeasibly demanding in practice, whether because A does not have sufficient internal capacity or A does not have appropriate external support. For example, expecting a human to take over operational control of an autonomous vehicle safely in under ten seconds is unfeasible [18]. Fourth, it should not conflict with other tasks or duties A bears. For example, an engineer's task role-responsibility to delivering the product on time should not conflict with their moral and legal duties to ensure that it is tolerably safe. We have outlined the same criteria for all role-responsibilities and all actors, even though AI-based systems can only be task role-responsible. The intention is that this general set of criteria provides a basis for readers to consider and discuss the clarity, appropriateness, demandingness and potential conflicts of role-responsibilities for actors and specific contexts.

There are connections between role-responsibility and both liability and moral responsibility. The possession of a legal duty is a requirement for A to be liable to legal sanction for O. Further, we have decomposed moral responsibility such that A's failing to do one's moral duty is a requirement for A to be liable to moral sanction for O [25]. Misunderstanding of the connection between role-responsibility and the other kinds could lead to unwittingly placing a disproportionate burden of legal or moral sanction on individual role-holders. Take, for example, the UNESCO Recommendation on the Ethics of Artificial Intelligence: *"The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system"* [106]. The 'always' here suggests that having a role-responsibility with respect to some part of the AI life cycle is both necessary and sufficient for liability and moral responsibility. This is misguided: role-responsibility is necessary but not sufficient for liability and moral responsibility. System outputs, or humans decisions and actions based on these outputs, may lead to harm without any of the role-holders having done something wrong, for example in the cases of emergent behaviour or unexpected occurrences in the operating environment. Getting clear on this can help us to ensure that designated role-holders are not immediately, and unfairly, held legally and morally responsible for all adverse outcomes.

2.2.3 Is legal-liability responsible for. As we continue to unravel the different senses in which an actor can be responsible for occurrences, we come to liability. An actor who acts unlawfully is usually liable, according to other legal rules, to sanction, whether punishment or the payment of compensation [52].

This paper divides legal liability-responsibility into criminal and civil liability only. These subtypes are shown in Figure 4. For present purposes, public law, which regulates the behaviour of public bodies, is omitted from this paper's analysis. The analysis focuses on the common law, the system of law which evolved in England and Wales. Common law legal systems include England and Wales, the United States, Australia, Ireland, New Zealand, Canada, India, Hong Kong, Singapore, Ghana, Uganda and Jamaica (amongst others). However, many legal principles found within the common law are also found across developed legal systems, for instance in Civil law systems, many having a common taxonomic or historical root. Because only natural and legal persons can be legally liable for occurrences, Figure 4 picks out just individual humans and institutions as the relevant categories of actor *A*. Further, because to trigger liability an (actionable) harm actually has to have happened, the only subcategory of occurrence in the Figure 4 is consequences (these are explained further in Section 2.3).

Criminal law concerns the prosecution of behaviour in the criminal courts. It primarily aims to safeguard the public and punish harmful acts, and is concerned with, *inter alia*, the protection of bodily integrity, of property and of the public welfare. Conditions for criminal liability are an action element (*actus reus*) – the committing of the conduct necessary for a crime – and a mental element (*mens rea*) – the doing so with the requisite intention, recklessness, or

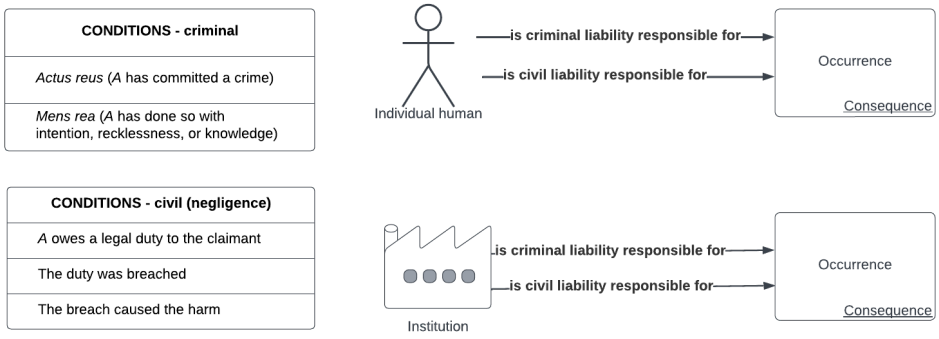

Fig. 4. Actor A is legal liability-responsible for Occurrence O

knowledge. Importantly, however, one kind of criminal liability that may prove salient in cases involving AI, namely strict criminal liability, which is often used in regulatory offences, lacks the *mens rea* element. Another is secondary liability, which considers the aiding or encouragement of the completion of a criminal offence by a principal offender.

Civil law regulates behaviour between parties. It seeks to determine the rights and duties of natural and legal persons, for example by establishing civil liability for a harm or 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). Conditions for civil liability also vary, so we pick out those in areas of civil liability that are particularly pertinent to cases involving AI. First, tort law (of which there are several subcategories). The workhorse tort in common law jurisdictions is negligence. Its elements, or conditions, are as follows: the actor owes a legal duty (of care) to the claimant; this duty was breached; the breach caused the claimant's harm. This example is given in Figure 4. Another type of tort is strict product liability (although defective products in some circumstances may also lead to liability in negligence). Vicarious liability, when an actor (typically an employer) is held civilly liable for the torts of another (typically their employee), is not a separate tort but a way in which any of the torts can be attributed. Second, in addition to tort law, contract law is a relevant branch of liability for AI-enabled systems, although contractual duties are only owed between contracting parties, and are not owed to third parties (except in very limited cases). Here, the condition is that an actor must have breached terms of their contract.

There are connections between liability, causal responsibility, and role-responsibility. When actors cause harm by breaching a legal duty they are, broadly speaking, liable for the consequences. The connection between legal responsibility and moral responsibility is a faultline debate in the philosophy of law. Legal positivists argue that, while the law and morality are often in agreement, what is required or allowed by law may be prohibited by morality, and vice versa [51, 53]. Legal moralists, by contrast, take the law to be an essentially moral enterprise, and thus reject the view that law and morality can be separated [29]. Speaking clearly about liability and moral responsibility does not require taking a side in this debate.

There is a specific concern that an individual actor, typically the immediate human-in-the-loop, can be unfairly attributed liability is by being a 'liability sink' [64]. Whether a clinician using an AI-based system or a remote operator in fallback control of multiple autonomous vehicles or ships, the worry here is that the individual will bear the sole burden of liability, where justice would require that it is distributed across a range of individual and institutional actors in the life cycle of the AI-based system [64, 99]. It arises because that individual is often the easiest actor to sue [64]. A real-world case of the liability sink is the sentence of endangerment of the safety driver in the Uber vehicle that

killed Elaine Herzberg in Arizona in 2018. Though the safety driver had certainly been acting negligently, she alone carried liability for the incident, because Uber ATG settled the civil case brought by the victim's family out of court (and prosecutors said that the corporation was not criminally liable). The corollary of the liability sink is that it can help to obscure the negligence of individual and institutional actors upstream in the life cycle. By focusing on the nearest individual in the scene, legal accountability for upstream omissions such as failing to build sufficient risk mitigation systems into the vehicle, failing to implement a good safety culture or employing engineers with inappropriate training, which contributed to the incident [7, 91], can be avoided.

2.2.4 Is morally responsible for. The fourth sense in which an A might be responsible for an O is moral responsibility. In everyday speech, people talk about their 'moral responsibilities', meaning their moral duties or obligations. These are covered in our vocabulary under role-responsibility. By 'moral responsibility', we refer to the relation between an actor *A* and an occurrence *O* that makes it appropriate to ask the actor *why* they acted as they did and react to A in ways such as resentment or gratitude, and corresponding practices such as blame and praise, depending on their reasons [100].

We draw a distinction between *being* morally responsible (moral attributability) [92, 116] and *being held* morally responsible (moral accountability) [116]. While there are other, more fine-grained taxonomies of responsibility in the philosophical literature [96], we restrict this paper's subcategories of moral responsibility to two for ease of use. These two subcategories of moral responsibility are shown in Figure 5. As discussed in Section 3, AI-based systems are not included amongst the subcategories of actor that can be morally responsible for an O, and they are therefore excluded from Figure 5.

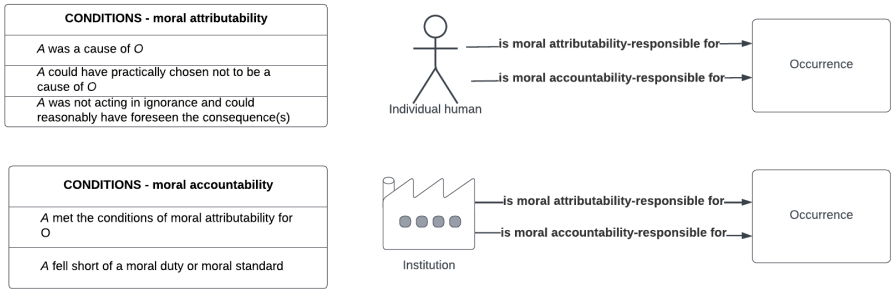

Fig. 5. Actor A is morally responsible for O

Moral attributability – being responsible – concerns an actor's relation *to their own* behaviour. An O is morally attributable to an A when A is the author of O. It is on the grounds of their authorship of O that A, who is a moral agent, would be rightly called upon to answer for O, whether this is an action, an omission, or a consequence. This opens A up to reactions, judgements and practices from the moral community (see moral accountability below). Judgements of blameworthiness do not necessarily follow from an A's authorship of O; it depends on the reasons for which A acted or failed to act. This is important to bear in mind for those inclined to shy away from talking about moral responsibility, for similar reasons to the reluctance to be explicit about causal responsibility - in case it leads unhelpfully to pointing the finger rather than avoiding harm. Moral responsibility as attributability requires an actor to explain their reasons; it is to be open to giving a particular kind of explanation, specifically a justification for their role in an AI-based system's outputs or impacts [85]. Not only does this not entail blame, it also helps to understand how to avoid future incidents and accidents, acknowledging this and learning from their reasons for action can help to improve processes and procedures.

Moral accountability – being held responsible – concerns an actor's relations *to other people*. A is morally accountable when they are subject to interpersonal responses such as blame and indignation, and practices like punishment and demands for apology, for their authorship of the O [25, 70, 100]. These practices are what we refer to in Table 1 as 'moral sanction'. Though it is not included in the conditions in Figure 5, another consideration is whether it is appropriate to *express* blaming reactions or similar towards A. For example, in an organisational culture that promotes openness and the confidence to raise concerns, it is not *always* appropriate actively to blame individuals for their mistakes, even when they've fallen below expectations [26]. But even if it is not always hold an actor morally accountable, that is not to say that their action, omission or its consequence is not morally attributable to them.

There is a presumed entailment relation between moral accountability and moral attributability [97, 116]; in other words, *being* morally responsible for O is a necessary condition for rightly *being held* morally responsible for O.

Distilling a rich history in the philosophical analysis of moral responsibility, and drawing some stipulations, we propose the following conditions for moral responsibility as attributability: A was a cause of, or causally responsible for, O; A's causing O was an exercise of their voluntary agency and they were not under undue pressure or duress (control condition); A was not acting in ignorance and could have reasonably foreseen the consequences (knowledge condition) [3, 38, 43, 117]. Failing to meet these conditions would excuse A from moral responsibility as attributability. We propose the following additional condition for moral accountability: A fell short of a moral duty or a standard of care or respect for others that was reasonably expected of them [25, 100]. There may reasons not to blame an A who fell short of a duty: perhaps it was unfeasibly demanding in the circumstances, as per the criteria set out in Section 2.2.2sec:). In cases of praise and gratitude, it would not be that A voluntarily fell short of a duty or an expected standard but that they met or exceeded such a standard. It is understandable, but perhaps regrettable, that the discussion of responsibility for AI focuses more on when it would be appropriate to blame rather than when it would be appropriate to praise AI actors. Indeed, a fruitful avenue of consideration could be 'praiseworthiness for AI', establishing what AI actors would be praiseworthy for [62, 83, 108]. However, in this discussion we restrict the condition to falling short of a moral duty, and subsequent liability to moral sanction.

Conceptual clarity about moral responsibility and its conditions can help us to be alert to possible unfair ascriptions of moral responsibility. Akin to the liability sink, the 'moral crumple zone' arises when an individual human actor carries the full burden of blame for the adverse consequences of an AI-based system's outputs or actions, typically due to a failure to intervene before the outputs of system cause harm [34]. Attending to the conditions of moral responsibility as attributability reveals why the moral crumple zone phenomenon is often unfair: because the individual in question did not meet the control condition sufficiently robustly and/or they may not adequately meet the knowledge condition. Of course, sometimes the individual in-the-loop does meet these conditions, but unfairness arises rather because there are other causally involved actors who meet the conditions of moral responsibility and the individual in-the-loop is absorbing all of this moral responsibility which should rightly be attributed to others as well. Though conceptual clarity alone will not mean that those other actors are placed under the spotlight and asked to justify the reasons for their actions which contributed to a given incident, accident or harm, it does at least help to remind the users of this work to think about which actors may be escaping moral censure in specific cases, such as those actors whose causal contribution upstream in the life cycle was an exercise of their voluntary agency with foreseeable, negative downstream consequences.

Another normative problem, which is exacerbated by *increasingly autonomous* AI-based systems, is the problem of moral responsibility gaps [46, 47, 55, 61, 69, 84, 105]. The problem here is that the immediately harm-causing entity, an AI-based system which can have real-world impacts without human intervention, is not a moral agent and hence

exempt from attributions of moral responsibility, while the relevant individual and institutional actors do not meet, to a sufficient degree, the control and knowledge conditions with respect to the AI-caused O and hence are *excused* from attributions of moral responsibility [47, 61, 84]. It has been acknowledged that a natural inclination when faced with autonomous systems is to argue that there needs to be a human-in-the-loop so that *someone* can be held responsible [23] – but, as we have seen above, that can be outright unfair to that individual if they do not meet the requisite control or knowledge conditions. Some scholars are sceptical that the responsibility gap problem even exists, arguing that senior actors such as military commanders do have sufficiently robust control [55] or that the machines themselves can be responsible in attenuated ways and that they can be 'sanctioned' in the form of control, training and management [105]. Another perspective is that the responsibility gap problem exists only in a minority of cases, specifically cases of unforeseen emergent behaviour from the AI-based system, and that even in these cases moral attributability traces back to the individual and institutional actors who voluntary committed to the risk of deploying the systems and that moral accountability turns on whether or not these actors fell short of the risk bearers' reasonable expectations of respect and regard [84].

2.3 Occurrence O

In the project of locating responsibility for AI, it is important to be clear and specific about *what*, exactly, different AI actors are purported to be responsible *for*. How an occurrence is described often influences where responsibility is located [104]. If we describe O in very general terms, for example the safety of the AI-enabled system, it will be more difficult to identify specific actors who are responsible for it. And an actor may be responsible for one aspect of the occurrence, for example acting on an AI radiography system's diagnosis, but not for another, for example the decision-threshold at which the model was trained to classify images as a high-risk concern. Specificity is critical to support fair and reasonable attributions of responsibility. It is also central to safety analysis practices, which focus on specific failure modes, or or ways in the system could fail, and the occurrences that could cause it [12, 37].

We propose seven subcategories of occurrence, which are presented in Figure 6.

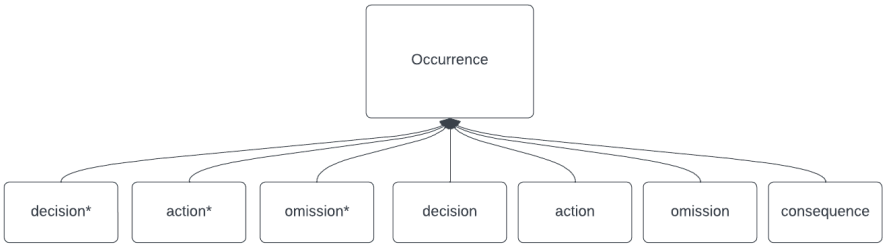

Fig. 6. Subcategories of occurrence *O*

First, there are the outputs of the AI-based systems. These outputs may be classifications, predictions, recommendations or generated text, speech or images. We refer to these as 'decisions*'. For physically-embodied or robotic AI-enabled systems, continuous sequences of physical manoeuvres would also count as their outputs. We refer to these outputs as 'actions*'. And we include 'omissions*' for cases where an AI system fails to generate an output. The star is included to distinguish the machine's outputs (or omissions) from a human's. This is useful to support diagrammatic reasoning about responsibility in complex cases involving AI-based systems; it helps to make clear that certain lines of

inquiry, such as inquiry into liability and moral responsibility must trace back to actors who build, use or regulate the systems.

Second, there are the decisions, actions and omissions of individual humans and institutions. Individual designers, data scientists, engineers, corporate managers, operators, users, regulators and legislators make legion decisions and take legion actions related to the design, maintenance and use AI-based technologies. They also make decisions about the organisational cultures in which the systems are developed, and criteria for system approval and risk acceptance. Institutions also make decisions and take actions in the form of board votes, corporate strategies and policies. And where it is unclear which individuals within an institution took decisions, actions or failed to act, it can be helpful in the first instance to attribute responsibility for these occurrences to institutions.

Third, there are consequences. Consequences are the tangible outcomes and experiences that are caused by any of the other subcategories of O as well as by physical events in the operating environment, such as storms and earthquakes. In inquiries of liability, it will be a consequence, such as a collision, injury, loss of life, damage to property or environmental damage for which liability will be sought. Consequences are not distinguished into those that are caused by individual humans and institutions (through their decisions, actions or omissions) and those that are caused directly by AI-based systems (by their decisions*, actions* or omissions*).

3 APPLYING THE CONCEPTUAL ANALYSIS AND GRAPHICAL NOTATION TO MODEL A RESPONSIBILITY SCENARIO

We have used the metaphor of unravelling responsibility to describe the conceptual breakdown of responsibility into different ways in which the statement 'Actor A is responsible for Occurrence O' could be understood. This was accompanied by an informal graphical notation to facilitate diagrammatic reasoning about responsibility for AI. Here, in Section 3, we illustrate the work by application to a hypothetical collision between involving an autonomous ship.

3.1 Example case: fatal collision between autonomous ship and traditional crewed vessel

Imagine an autonomous, crewless ship. For the most part, it is controlled by a human operator from a shore-based remote operation centre (ROC). But the ship has the capability to operate autonomously, without human intervention, and can do so where agreed with the regulatory authorities. The ship and the whole maritime autonomous infrastructure are supposed to be safe by design and, where this is not possible, fault-tolerant.

This ship is transporting cargo between Port A and Port B. It is being remotely controlled from the ROC by the remote human operator. The remote operator sees that, if the ship does not alter its course, there will be a collision with another vessel. The second vessel is a traditional boat crewed with STCW (Standards of Training, Certification and Watchkeeping for Seafarers) personnel. Before the remote operator can take action to adjust the ship's speed or course to avoid the collision, as required by the Convention on the International Regulations for Preventing Collisions at Sea (COLREGS), there is a loss of connectivity between the ROC and the autonomous ship. The remote operator therefore cannot send a control signal to the autonomous ship.

The ship detects the loss of connectivity and shifts to autonomous operation. The ship is fitted with an ML-based collision-avoidance system, which relies on sensors on the vessel, including cameras and lidar, to identify other vessels in the area and take avoidant action when required. It is raining heavily and the camera lenses are covered in salt. This degrades the cameras and the ship's sensors do not detect the other vessel. As a consequences, the ML-based collision-avoidance system does not properly work. The crew on board the other vessel expect the autonomous ship to comply with the International Maritime Organisation's (IMO) regulations for preventing collisions at sea. It does not

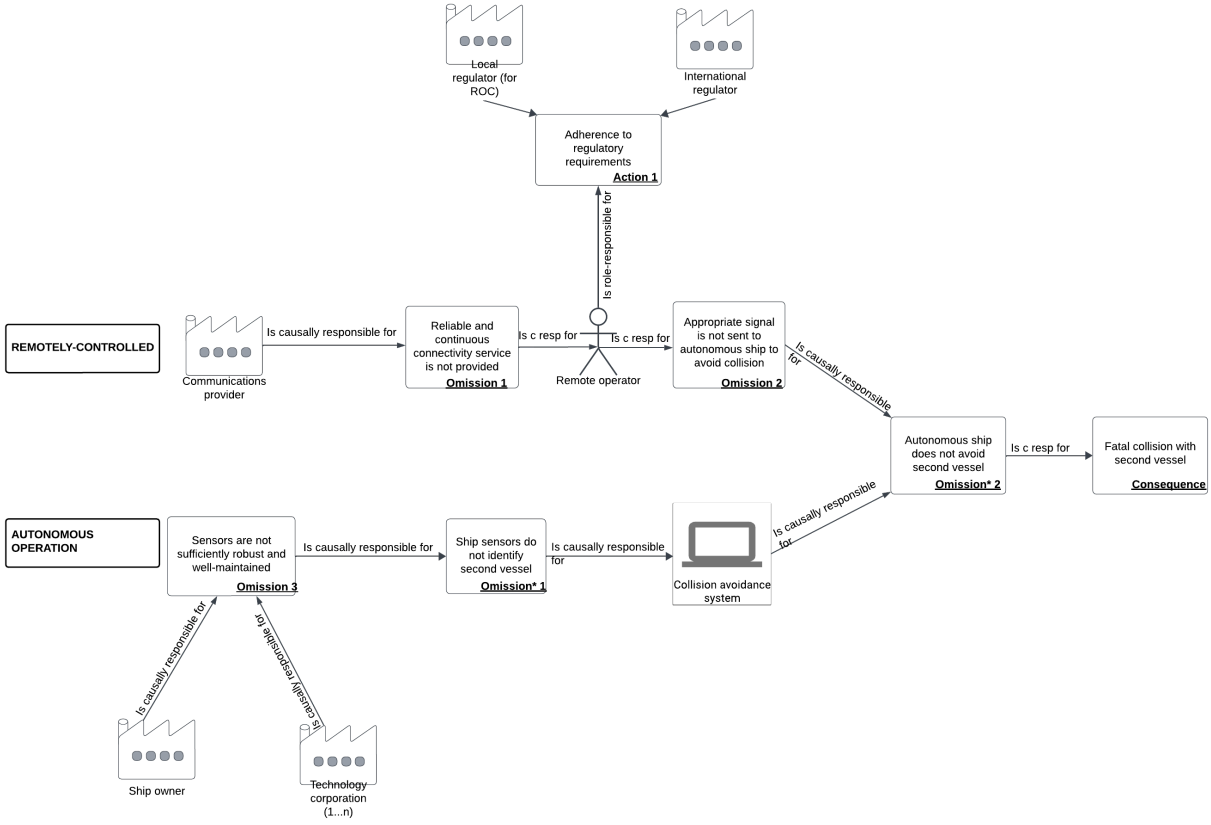

Fig. 7. First stages of modelling responsibility for AI in the maritime example, with arrows primarily denoting causal responsibility

and, because the autonomous ship is both crewless and has lost connectivity, there is no way for the other vessel to communicate with it. The other vessel does not have enough time to take evasive action. The collision occurs, causing the loss of life of several members of the other vessel's crew. Significant environmental damage is also caused, due to a large spillage of oil into the sea.

3.2 Modelling the scenario

Figure 7 is a partial model, using the informal notation introduced in in Section 2, which shows how the first stages of mapping this scenario might proceed. It only depicts the initial stage of modelling, in which the *causally responsible* actors and occurrences are set out. As discussed in Section 2.2.1, being a causal factor in the occurrence is necessary for an actor *even to be considered* for attributions of other senses of responsibility. Building up from causal responsibility may therefore a helpful way to *start* reasoning about responsibility for AI. The central scenario depicted in Figure 7 is that the remote operator would have made adjustments to the ship's speed or course to avoid the collision, but the loss of connectivity made this impossible. The ship therefore switched to autonomous mode but the degradation of the cameras meant that its collision-avoidance system did not work properly, and the fatal collision ensued.

As a rule of thumb, the following methodology was followed. This is one way in which the conceptual analysis and informal graphical notation could be used to model scenarios involving AI and where we seek to understand who is responsible, in what sense, and for what. Other approaches, such as starting with role-responsibility, could be developed [90]. All occurrences referred to are as labelled in Figure 7.

- First, we identified the main occurrence for which responsibility is sought. To manage complexity, we focused on fatal collision (Consequence in Figure 7) and not on the environmental damage.
- Second, we traced back from the collision to occurrences that caused it. As a matter of judgement, some causal factors, such as the weather, were not included in the model. The occurrences meet the NESS threshold for causality outlined in Section 2.2.1. For example, the absence of a command to the autonomous ship from the human remote operator (Omission 2) is a necessary element of a set sufficient for the fatal collision. Other necessary elements of the set include Omission 1, Omission 3 and Omission* 1. The numbering of the occurrences shown as rectangles in Figure 7 is not to denote any priority of importance or weight, but is simply to facilitate systematic discussions.
- Third, we traced back from those occurrences to actors who were causally responsible for them. The actors identified are also a partial selection in the model. The AI-based actor is collision-avoidance system. The only individual human actor specifically depicted is the remote human operator. The institutions cover the communications provider, local and international maritime regulators, the ship owner, and the technology corporation(s) who developed different systems in the ship.
- Fourth, though the identification of different role-responsibilities is by no means complete, we also identified a role-responsibility that remote operator has to adhere to regulatory requirements (Action 1). This would be a legal duty.

3.3 Reasoning about causal, role, liability and moral responsibility for AI in the scenario

The aim of the example is to show how the unravelling of the concept of responsibility and accompanying notation could be used to model complex events for which responsibility is sought. The approach taken here is start with causal responsibility – which in this paper is another term for making a causal contribution to an occurrence – as a general entry point into the discussions, as represented in Figure 7, and then to layer up from there, adding role, liability and moral responsibility, as well as additional occurrences and actors, as discussions progress. These would supply additional lines and elements into the model, which can be iteratively updated. As mentioned above, users of the work may wish to apply the notation differently to reason about responsibility for AI in specific cases.

3.3.1 Causal responsibility. Figure 7 already picks out relations of causal responsibility between actors (and occurrences) and other occurrences in the scenario. But one way in which the approach could support more in-depth discussions about causal responsibility for AI is as a basis for considering the relative weight of different actors' causal contributions. For simplicity and maximum generality, in Section 2.2.1 we proposed the NESS condition as a threshold condition for causality. But that does not mean that further conversations cannot be had about whether a scalar notion of causality could be appropriate. Several different measures for degrees of causal weight have been proposed in the literature [8, 27, 60, 71, 102]. These include whether the cause constituted a greater proportion of the outcome, such as the collision, or a greater proportion of the cause, such as the relative contribution it might have made Omission* 2 in Figure 7 [102]. Focused conversations about whether different degrees of causal responsibility should influence degrees of attribution of responsibility in the other senses could then ensue [60, 102].

3.3.2 Role-responsibility. Discussants might consider which actors had a role-responsibility to maintain safety. With respect to the remote operation context, we should consider the tasks and duties of the communications provider (institution), the individual remote operator in the ROC (individual human), and regulators (institutions) issuing guidelines. With respect to the autonomous ship itself (AI-based system), we should consider the tasks and duties of the ship's maintenance team (individual humans) and system developers (institutions). One way in which discussants could now reason about role-responsibility is to cross-reference against the criteria given in section 2.2.2, and consider feasible tasks and duties for each of the identified actors in Figure 7, as well as for actors newly identified. Heavy demands on individuals 'in-the-loop' as well as conflicts between roles and duties could be made explicit [2, 23, 35], to preempt and avoid later inappropriate or unfair attributions of role-responsibility.

3.3.3 Liability. There are several interests at stake (property interests, insurer interests, cargo interests, and interests connected to the environmental damage) for which claims could be made, along with potential criminal prosecution for pollution offences. But for present purposes, we focus on the deaths of the seafarers. To unravel civil liability for the deaths, it needs first to be established where legal duties lie, and whether any of these were breached. For example, were the ship owners (of which there are, in practice, multiple layers of institutions) meeting their duty to ensure adequate inspection, maintenance and concept of operation of the cameras (Omission 3)? Did the corporation who installed them do so correctly, or provide proper instruction as to their maintenance (Omission 3)? Did the corporation who designed and manufactured the cameras consider the maritime environment, where they would be regularly exposed to salt water, and salt deposits (Omission 3)? Did the communications provider meet its duties of adequately maintaining the service, or supplying appropriate hardware or a service of satisfactory quality (Omission 1)?

From these questions, we can also consider possible criminal liability. If the duties of care have been breached, one or more of these actors could face criminal prosecution, for instance via a charge of gross negligence manslaughter or corporate manslaughter (in the case of institutions), provided that the nature of the defendant's breach of its duty of care owed to the victim amounts to a gross breach of its duty. For these two offences the *mens rea* element is typically considered to be a high degree of negligence. In addition there may be offences committed under the Merchant Shipping Act 1995, for instance the potential for owner liability under Section 98 (dangerously unsafe ship); or the potential for owner, charterer, or manager liability under Section 100 (unsafe operation of ship). Further, the ship's failure to comply with COLREGS may result in the owner or any person 'responsible' for the conduct of the ship, being criminally liable under Regulation 6 of The Merchant Shipping (Distress Signals and Prevention of Collisions) Regulations 1996, unless they can show that they 'took all reasonable precautions to avoid the commission of the offence.' The latter is a strict liability regulatory offence, with no *mens rea* requirement, but subject to a defence.

3.3.4 Moral responsibility. Discussants might consider whether the collision, or the omissions leading up to it, were things over which actors had voluntary control, or could practically choose whether or not to cause, and whether could reasonably have foreseen the consequences. For example, it seems clear that the remote operator's failure to send a control signal to the autonomous ship (Omission 2) was non-voluntary. As such, the remote operator cannot be said to be morally responsible for Omission 2. We might, however, question whether there was anything else the remote operator could have done, such as raise an alarm, which might have prompted actions which would have ensured that Omission*2 did not occur, or did not cause the fatal collision. But *prima facie*, it does not seem that the remote operator is rightly morally responsible for the fatal collision. Clearly, improperly maintained sensors are serious safety hazards that could foreseeably render a collision-avoidance system ineffective, and this was something which both the relevant individual actors and the relevant corporations or institutions could control. But inquiries into

moral responsibility for the collision would consider whether inadequate maintenance of the cameras was due to a lackadaisical attitude amongst some actors or whether they were they under significant pressure to save time and cost? Did powerful institutional decision-makers in the scenario adequately put in place the structures and cultures to make Omission* 2 less likely? As we have seen, it is sadly likely that individuals close to the occurrences, the last 'lines of defence', will end up being subject to disproportionate moral blame and sanction when things go wrong. Thinking about whether causally connected and voluntary decision-makers upstream fell short of a moral duty or an expected standard of regard for the risk-exposed can help us to highlight where we may need to address asymmetries between the weight of moral accountability borne by individuals and big players.

4 DISCUSSION

In our experience, there is considerable confusion about the meaning of the term responsibility. This confusion stands in the way of precise and targeted deliberations about where responsibility lies for the outputs and impacts of AI-based systems. Acknowledging that responsibility is essentially relational – that it describes a particular kind of relation between an actor and an occurrence – we have unravelled the concept of responsibility by decomposing the three-part statement 'Actor A is responsible for Occurrence O' to elucidate that there are different possibilities of *who* (or what) is responsible for the outputs and impacts of AI-based systems, the *sense* or *way* in which they are responsible, and *aspects of events* they are responsible for. The aim has been to clarify the concept of responsibility for AI for policymakers, practitioners, researchers and students from non-philosophical and non-legal backgrounds who are concerned with the moral and legal implications of the deployment of these technologies. To be clear, this work has not proposed solutions to problems such as responsibility gaps and the problem of many hands, which the use of AI-based systems introduce or exacerbate. Rather, it is a precursor to such problem-solving, with the aim of widening participation in these debates.

This conceptual analysis has been supplemented by an explicit statement of the criteria and conditions for fitting attributions of responsibility in each of the four senses: causal responsibility, role-responsibility, liability responsibility and moral responsibility. Clarity on these senses and their criteria or condition can highlight, and help to forestall, two undesirable situations. First, actors trading on ambiguity or using equivocal language about responsibility to avoid taking responsibility for the outputs and impacts of AI-powered systems. Second, individual actors disproportionately absorbing the blame and liability that should be shared with others – a phenomenon that has come to be known as the 'moral crumple zone' or 'liability sink' problem [34, 64]. We have distilled central distinctions in philosophy and law to present this work. It is hoped that by making this explicit, those who disagree with this paper's analysis of responsibility can easily pinpoint where and why and, if desired, propose and justify changes. We have also presented an informal graphical notation and illustration of how it can be used to model specific cases involving AI-based systems.

There are several possible uses of the work. One intended audience is policymakers and regulatory officials in the different sectors in which AI is being deployed. The paper may be used by these stakeholders by providing a basis for extending questions of 'who' is responsible for the outputs and impacts of AI-based systems to 'what for?' and 'how?'. In respect of role-responsibility, UK sector-specific regulators are tasked with translating high-level principles for responsible AI into concrete guidance for AI actors across the life cycle. Thinking about the criteria for appropriate attributions of role-responsibility, and what these actors are responsible for, can help to ensure that tasks and duties are appropriately targeted and feasible [42]. While the EU's approach, as set out in the AI Act, is more prescriptive, mandating specific legal duties around conformity assessment, auditing and the monitoring of high-risk systems, the clarification offered by this paper can facilitate critical thinking about possible conflicts of an actor's legal duties and their other role-responsibilities [16]. It may be similarly helpful to policy practitioners in the US, where the Executive

Order 14110 on Safe, Secure and Trustworthy Artificial Intelligence puts some legal duties directly on AI developers, such as sharing safety test results with the US government, and some duties on bodies such as the National Institute of Standards and Technology (NIST), such as the development of AI-related risk management standards. The NIST's draft Risk Management Framework for Generative AI sets out tables of actions to manage risks from Generative AI, and states clearly that not all actions apply to all AI actors [77]. One way to evaluate the viability of these actions would be to consider whether they meet the criteria for appropriate attributions of role-responsibility set out in this paper, including internal conflicts within them and where they might conflict with other task responsibilities.

It is not just in respect of role-responsibility that public officials could find value in this work. Being cognisant of the ways in which different senses of responsibility, and the relations between them, can be confused can help them to be alert to and to forestall evasions of responsibility, for example when developers or operators speak of transferring moral duties or moral responsibility to increasingly autonomous systems, or the scapegoating of individuals, for example when the nearest causally involved human is disproportionately subject to legal and moral sanction.

Another intended audience is practitioners, particularly safety engineers, and technical researchers. Safety engineering practices involve predicting what events will cause or raise the probability of hazards which endanger human life or damage property or the environment. These hazards are modelled using methods such as fault trees to identify which potential failures or faults need to be eliminated or mitigated, and to enable assessment of the sufficiency of the mitigations with regard to acceptable levels of hazard risk[19]. The conceptual analysis and notation presented is not intended to replace such activities, but may offer a useful perspective that can enrich and help to guide safety practices [90]. It can, for example, help practitioners to think critically about whether the role-responsibilities upon operators to intervene are feasible, and to consider new forms of endangerment such as individuals bearing disproportionate liability or moral responsibility for outcomes over which they have limited control or knowledge.

We also envisage that the paper has value for students from diverse disciplines who are interested in the attribution of responsibility for AI. By providing conceptual clarity on what 'responsibility' means, by summarising a body of literature, by providing a distilled set of criteria and conditions for appropriate and fair attributions of responsibility, and then presenting these in an informal graphical notation, it can help students to engage in more sophisticated debates about where responsibility does and should lie for the outputs and impacts of AI in specific cases.

5 CONCLUSION

This paper has presented a systematic unravelling of the concept of responsibility, and has showed how it can be used to model scenarios and reason about responsibility for the outputs and impacts of AI-powered technologies in a focused way. Its purpose has been to furnish clarity and specificity, at an accessible level of abstraction, to facilitate transparent discussions about fair and fitting attributions of responsibility for AI.

ACKNOWLEDGMENTS

The work was supported by the Engineering and Physical Sciences Research Council (EP/W011239/1). The authors thank Professor Paul Noordhof, Professor Ian Pyle and other reviewers for their comments on the work.

REFERENCES

[1] Amanda Askell, Miles Brundage, and Gillian Hadfield. 2019. The role of cooperation in responsible AI development. *arXiv preprint arXiv:1907.04534* (2019).

[2] Lisanne Bainbridge. 1983. Ironies of automation. In *Analysis, design and evaluation of man–machine systems*. Elsevier, 129–135.

[3] Robert C Bartlett, Susan D Collins, et al. 2011. *Aristotle's Nicomachean ethics*. University of Chicago Press.

[4] Helen Beebee, Christopher Hitchcock, Peter Charles Menzies, and Peter Menzies. 2009. *The Oxford handbook of causation*. Oxford Handbooks.

[5] Reuben Binns. 2018. Algorithmic accountability and public reason. *Philosophy & technology* 31, 4 (2018), 543–556.

[6] Gunnar Björnsson and Kendy Hess. 2017. Corporate Crocodile Tears? *Philosophy and Phenomenological Research* 94, 2 (2017), 273–298.

[7] 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

[8] Matthew Braham and Martin Van Hees. 2009. Degrees of causation. *Erkenntnis* 71 (2009), 323–344.

[9] Matthew Braham and Martin Van Hees. 2012. An anatomy of moral responsibility. *Mind* 121, 483 (2012), 601–634.

[10] Michael E Bratman. 2013. *Shared agency: A planning theory of acting together*. Oxford University Press.

[11] Peter Cane. 2002. *Responsibility in law and morality*. Bloomsbury Publishing.

[12] Carl S Carlson. 2012. *Effective FMEAs: Achieving safe, reliable, and economical products and processes using failure mode and effects analysis*. Vol. 1. John Wiley & Sons.

[13] Simon Chesterman. 2020. Artificial intelligence and the limits of legal personality. *International & Comparative Law Quarterly* 69, 4 (2020), 819–844.

[14] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1186–1197.

[15] Mark Coeckelbergh. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics* 26, 4 (2020), 2051–2068.

[16] European Commission. 2021. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[17] European Commission. 2022. Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[18] Law Commission and Scottish Law Commission. 26 January 2022. Automated Vehicles: summary of joint report, HC1068. https://lawcom.gov.uk/project/automated-vehicles/

[19] US Regulatory Commission. 1981. Fault Tree Handbook NUREG-0492. https://www.nrc.gov/docs/ML1007/ML100780465.pdf

[20] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 864–876.

[21] David Copp. 1979. Collective actions and secondary actions. *American Philosophical Quarterly* 16, 3 (1979), 177–186.

[22] Roger Crisp. 2014. *Aristotle: nicomachean ethics*. Cambridge University Press.

[23] Rebecca Crootof, Margot E Kaminski, W Price, and II Nicholson. 2023. Humans in the Loop. *Vand. L. Rev.* 76 (2023), 429.

[24] David Danks, David Rose, and Edouard Machery. 2014. Demoralizing causation. *Philosophical Studies* 171 (2014), 251–277.

[25] Stephen Darwall. 2007. Moral obligation and accountability. *Oxford studies in metaethics* 2 (2007), 111–132.

[26] Sidney WA Dekker and Hugh Breakey. 2016. 'Just culture:'Improving safety by achieving substantive, procedural and restorative justice. *Safety science* 85 (2016), 187–193.

[27] Huzeyfe Demirtas. 2022. Causation comes in degrees. *Synthese* 200, 2 (2022), 64.

[28] Susannah Kate Devitt. 2022. Bad, mad, and cooked: Moral responsibility for civilian harms in human-AI military teams. *arXiv preprint arXiv:2211.06326* (2022).

[29] Patrick Baron Devlin. 1965. *Morals and the criminal law*. Oxford University Press Oxford.

[30] Online Etymology Dictionary. [n. d.]. responsible. https://www.etymonline.com/word/responsible

[31] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 2156. Springer.

[32] Robin Antony Duff. 2007. *Answering for crime: Responsibility and liability in the criminal law*. Bloomsbury Publishing.

[33] Damian Okaibedi Eke, Kutoma Wakunuma, and Simisola Akintoye. 2023. Introducing Responsible AI in Africa. In *Responsible AI in Africa: Challenges and Opportunities*. Springer International Publishing Cham, 1–11.

[34] Madeleine Clare Elish. 2019. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)* (2019).

[35] Mica R Endsley. 2023. Ironies of artificial intelligence. *Ergonomics* (2023), 1–13.

[36] NHS England. version 1, August 2022. Patient Safety Incident Investigation. https://www.england.nhs.uk/wp-content/uploads/2022/08/B1465-PSII-overview-v1-FINAL.pdf

[37] Peter Fenelon and John A McDermid. 1993. An integrated tool set for software safety analysis. *Journal of Systems and Software* 21, 3 (1993), 279–290.

[38] John Martin Fischer and Mark Ravizza. 1998. *Responsibility and control: A theory of moral responsibility*. Cambridge university press.

[39] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020).

[40] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14 (2004), 349–379.

[41] Philippa Foot. 1967. The problem of abortion and the doctrine of the double effect. (1967).

[42] Department for Science Innovation and Technology. March 2023. A pro-innovation approach to AI regulation, CP815. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

[43] Harry Frankfurt. 2018. Alternate possibilities and moral responsibility. In *Moral Responsibility and Alternative Possibilities*. Routledge, 17–25.

[44] Peter A French. 1984. *Collective and corporate responsibility*. Columbia University Press.

[45] Daniela Glavaničová and Matteo Pascucci. 2022. Vicarious liability: a solution to a problem of AI responsibility? *Ethics and Information Technology* 24, 3 (2022), 28.

[46] Trystan S Goetze. 2022. Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 390–400.

[47] Raul Hakli and Pekka Mäkelä. 2019. Moral responsibility of robots and hybrid agents. *The Monist* 102, 2 (2019), 259–275.

[48] Jaana Hallamaa and Taina Kalliokoski. 2020. How AI systems challenge the conditions of moral agency?. In *International Conference on Human-Computer Interaction*. Springer, 54–64.

[49] Sven Ove Hansson. 2022. Responsibility in road traffic. In *The Vision Zero Handbook: Theory, Technology and Management for a Zero Casualty Policy*. Springer, 1–27.

[50] Sven Ove Hansson. 2022. Who Is Responsible If the Car Itself Is Driving? *Test-Driving the Future: Autonomous Vehicles and the Ethics of Technological Change* (2022), 43.

[51] HLA Hart. 1958. POSITIVISM AND THE SEPARATION OF LAW AND MORALS t. *Harvard Law Review* 71, 4 (1958), 593–629.

[52] Herbert Lionel Adolphus Hart. 2008. *Punishment and responsibility: Essays in the philosophy of law*. Oxford University Press.

[53] Herbert Lionel Adolphus Hart and Leslie Green. 2012. *The concept of law*. oxford university press.

[54] Herbert Lionel Adolphus Hart and Tony Honoré. 1985. *Causation in the Law*. OUP Oxford.

[55] Johannes Himmelreich. 2019. Responsibility for killer robots. *Ethical Theory and Moral Practice* 22, 3 (2019), 731–747.

[56] Johannes Himmelreich and Sebastian Köhler. 2022. Responsible aI through conceptual engineering. *Philosophy & Technology* 35, 3 (2022), 60.

[57] The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[58] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[59] Deborah G Johnson. 2006. Computer systems: Moral entities but not moral agents. *Ethics and information technology* 8 (2006), 195–204.

[60] Alex Kaiserman. 2018. 'More of a cause': Recent work on degrees of causation and responsibility. *Philosophy Compass* 13, 7 (2018), e12498.

[61] Maximilian Kiener. 2022. Can we Bridge AI's responsibility gap at Will? *Ethical theory and moral practice* 25, 4 (2022), 575–593.

[62] Zoe Johnson King. 2023. What are we Praiseworthy For? In *Conversations in Philosophy, Law, and Politics, eds. Ruth Change and Amia Srinivasan*. Oxford University Press.

[63] Joshua Knobe. 2009. Folk judgments of causation. *Studies in History and Philosophy of Science Part A* 40, 2 (2009), 238–242.

[64] Tom Lawton, Phillip Morgan, Zoe Porter, Shireen Hickey, Alice Cunningham, Nathan Hughes, Ioanna Iacovides, Yan Jia, Vishal Sharma, and Ibrahim Habli. 2024. Clinicians risk becoming 'liability sinks' for artificial intelligence. *Future Healthcare Journal* 11, 1 (2024).

[65] Christian List. 2021. Group agency and artificial intelligence. *Philosophy & technology* 34, 4 (2021), 1213–1242.

[66] Christian List and Philip Pettit. 2011. *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

[67] John Randolph Lucas. 1993. *Responsibility*. Oxford University Press.

[68] Kirk Ludwig. 2017. *From plural to institutional agency: collective action II*. Oxford University Press.

[69] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6 (2004), 175–183.

[70] Michael McKenna. 2012. *Conversation & responsibility*. Oup Usa.

[71] Michael S Moore. 2009. *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford University Press, USA.

[72] Phillip Morgan. 2023. *Tort Liability and Autonomous Systems Accidents: Common and Civil Law Perspectives*. Edward Elgar Publishing.

[73] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2 (1996), 25–42.

[74] Paul Noordhof. 2020. *A variety of causes*. Oxford University Press.

[75] Sven Nyholm. 2023. Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In *Risk and responsibility in context*. Routledge, 191–213.

[76] OECD. 2019. AI principles: Accountability (Principle 1.5). https://oecd.ai/en/dashboards/ai-principles/P9

[77] National Institute of Standards and U.S. Department of Commerce Technology. April 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI-600-1, Initial Public Draft. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

[78] Chinasa T Okolo, Kehinde Aruleba, and George Obaido. 2023. Responsible AI in Africa—Challenges and Opportunities. *Responsible AI in Africa: Challenges and Opportunities* (2023), 35–64.

[79] Oxford English Dictionary online. [n. d.]. responsibility. https://www.oed.com/search/dictionary/?scope=Entries&q=responsibility

[80] Richard Owen, Jack Stilgoe, Phil Macnaghten, Mike Gorman, Erik Fisher, and Dave Guston. 2013. A framework for responsible innovation. *Responsible innovation: managing the responsible emergence of science and innovation in society* (2013), 27–50.

[81] Rodney Algernon Percy and John Charlesworth. 2018. *Charlesworth & Percy on Negligence*. Sweet & Maxwell.

[82] Philip Pettit. 2007. Responsibility incorporated. *Ethics* 117, 2 (2007), 171–201.

[83] Sebastian Porsdam Mann, Brian D Earp, Sven Nyholm, John Danaher, Nikolaj Møller, Hilary Bowman-Smart, Joshua Hatherley, Julian Koplin, Monika Plozza, Daniel Rodger, et al. 2023. Generative AI entails a credit–blame asymmetry. *Nature Machine Intelligence* (2023), 1–4.

[84] Zoe Porter. 2021. *Moral responsibility for unforeseen harms caused by autonomous systems.* Ph. D. Dissertation. University of York.

[85] Zoe Porter, Annette Zimmermann, Phillip Morgan, John McDermid, Tom Lawton, and Ibrahim Habli. 2022. Distinguishing two features of accountability for AI technologies. *Nature Machine Intelligence* 4, 9 (2022), 734–736.

[86] Thomas M Powers. 2013. On the moral agency of computers. *Topoi* 32 (2013), 227–236.

[87] Kostina Prifti, Evert Stamhuis, and Klaus Heine. 2022. Digging into the Accountability Gap: Operator's Civil Liability in Healthcare AI-systems. In *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice.* Springer, 279–295.

[88] Duncan Purves, Ryan Jenkins, and Bradley J Strawser. 2015. Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice* 18 (2015), 851–872.

[89] Alan Rubel, Clinton Castro, and Adam Pham. 2019. Agency laundering and information technologies. *Ethical Theory and Moral Practice* 22 (2019), 1017–1041.

[90] Philippa Ryan, Ibrahim Habli, Joanna Al-Qaddoumi, Zoe Porter, and John McDermid. 2024. What's my role? Modelling responsibility for AI-based safety-critical systems. (2024). https://arxiv.org/abs/2401.09459

[91] Philippa Ryan Conmy, John Mcdermid, Ibrahim Habli, and Zoe Porter. 2023. Safety engineering, role responsibility and lessons from the Uber ATG Tempe Accident. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems.* 1–10.

[92] Thomas M Scanlon. 2000. *What we owe to each other.* Harvard University Press.

[93] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2020. Principles to practices for responsible AI: closing the gap. *arXiv preprint arXiv:2006.04707* (2020).

[94] Miguel Ángel Sebastián. 2021. First-person representations and responsible agency in AI. *Synthese* 199, 3 (2021), 7061–7079.

[95] Amy J Sepinwall. 2016. Corporate moral responsibility. *Philosophy Compass* 11, 1 (2016), 3–13.

[96] David Shoemaker. 2011. Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics* 121, 3 (2011), 602–632.

[97] David Shoemaker. 2013. On criminal and moral responsibility. *Oxford studies in normative ethics* 3 (2013), 154–178.

[98] David Shoemaker. 2015. *Responsibility from the Margins.* Oxford University Press, USA.

[99] Helen Smith. 2021. Artificial Intelligence to Inform Clinical Decision Making: A Practical Solution to An Ethical And Legal Challenge. In *IJCAI 2020 AI for Social Good workshop.*

[100] Peter Frederick Strawson. 2008. *Freedom and resentment and other essays.* Routledge.

[101] Michael T Stuart and Markus Kneer. 2021. Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[102] Victor Tadros. 2018. Causal contributions and liability. *Ethics* 128, 2 (2018), 402–431.

[103] Dennis F Thompson. 1980. Moral responsibility of public officials: The problem of many hands. *American Political Science Review* 74, 4 (1980), 905–916.

[104] Dennis F Thompson. 2017. Designing responsibility: the problem of many hands in complex organizations. *Designing in ethics* (2017), 32–56.

[105] Daniel W Tigard. 2021. There is no techno-responsibility gap. *Philosophy & Technology* 34, 3 (2021), 589–607.

[106] UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence, SHS/BIO/PI/2021/1. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[107] Stanford University. 2023. AI Index Report 2023. https://aiindex.stanford.edu/report/

[108] Charlotte Franziska Unruh. 2022. Doing and allowing good. *Analysis* 82, 4 (2022), 630–637.

[109] Shannon Vallor and Bhargavi Ganesh. 2023. Artificial intelligence and the imperative of responsibility: Reconceiving AI governance as social care. In *The Routledge Handbook of Philosophy of Responsibility.* Routledge, 395–406.

[110] Ibo Van de Poel. 2011. The relation between forward-looking and backward-looking responsibility. In *Moral responsibility: Beyond free will and determinism.* Springer, 37–52.

[111] Ibo Van de Poel, Lambèr Royakkers, and Sjoerd D Zwart. 2015. *Moral responsibility and the problem of many hands.* Routledge.

[112] Carissa Véliz. 2021. Moral zombies: why algorithms are not moral agents. *AI & society* 36 (2021), 487–497.

[113] Mario Verdicchio and Andrea Perin. 2022. When doctors and AI interact: on human responsibility for artificial risks. *Philosophy & Technology* 35, 1 (2022), 11.

[114] Rene Von Schomberg. 2013. A vision of responsible research and innovation. *Responsible innovation: Managing the responsible emergence of science and innovation in society* (2013), 51–74.

[115] Wendell Wallach and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong.* Oxford University Press.

[116] Gary Watson. 1996. Two faces of responsibility. *Philosophical topics* 24, 2 (1996), 227–248.

[117] Jan Willem Wieland and Philip Robichaud. 2017. *Responsibility: The epistemic condition.* Oxford University Press.

[118] Ludwig Wittgenstein. 1976. Cause and effect: Intuitive awareness. *Philosophia* 6, 3-4 (1976), 409–425.

[119] Richard W Wright. 1985. Causation in tort law. *Calif. L. Rev.* 73 (1985), 1735.

[120] John Zerilli. 2021. *A citizen's guide to artificial intelligence.* MIT Press.