AI and Ethics
https://doi.org/10.1007/s43681-024-00549-9

ORIGINAL RESEARCH

Ethical risk for AI

David M. Douglas[1] · Justine Lacey[2] · David Howard[3]

Received: 30 May 2024 / Accepted: 4 August 2024

Abstract
The term 'ethical risk' often appears in discussions about the responsible development and deployment of artificial intelligence (AI). However, ethical risk remains inconsistently defined in this context, obscuring what distinguishes it from other forms of risk, such as social, reputational or legal risk, for example. In this paper we present a definition of ethical risk for AI as being any risk associated with an AI that may cause stakeholders to fail one or more of their ethical responsibilities towards other stakeholders. To support our definition, we describe how stakeholders have role responsibilities that follow from their relationship with the AI, and that these responsibilities are towards other stakeholders associated with the AI. We discuss how stakeholders may differ in their ability to make decisions about an AI, their exposure to risk, and whether they or others may benefit from these risks. Stakeholders without the ability to make decisions about the risks associated with an AI and how it is used are dependent on other stakeholders with this ability. This relationship places those who depend on decision-making stakeholders at ethical risk of being dominated by them. The decision-making stakeholder is ethically responsible for the risks their decisions about the AI impose on those affected by them. We illustrate our account of ethical risk for AI with two examples: AI-designed attachments for surgical robots that are optimised for treating specific patients, and self-driving 'robotaxis' that carry passengers on public roads.

Keywords Artificial intelligence · Ethical risk · Ethical responsibility · Stakeholders · Dependency

1 Introduction

As artificial intelligence (AI) and machine learning (ML) applications have become widespread, the risks associated with these systems have also become a topic of widespread interest. These risks range from biased decisions that reflect and reinforce existing social, racial, and gender inequalities [1–3] to failures in autonomous vehicles that lead to fatal accidents [4, 5]. These risks, both realised and potential, have inspired considerable interest in AI ethics, and how best to ensure that AI systems are designed and used in ways that reduce, mitigate, or avoid such harms. Part of this response is how to effectively address the *ethical risks* of AI systems [6].

In this paper we present a new account of ethical risk for AI. We argue that an *ethical risk for an AI system* is any risk associated with it that may cause stakeholders in the system to fail one or more of their ethical responsibilities towards other stakeholders. By 'stakeholders', we refer to the human agents (or groups of human agents) who may affect an AI system, or be affected by how others use it. Further to this, a stakeholder can also be *at ethical risk* from an AI system if they are dependent on another stakeholder who makes decisions about some characteristic of that AI system that may affect them in a way that means they can be wronged or harmed by the decision-maker's failure to fulfil their ethical responsibilities towards them.

To support this account, we bring together several concepts from the philosophy of technology, ethical responsibility, ethics of risk, and republican political theory that, to our knowledge, have not been combined into a single account of

✉ David M. Douglas
david.douglas@csiro.au

Justine Lacey
Justine.Lacey@csiro.au

David Howard
David.Howard@data61.csiro.au

[1] Socio-Technical Innovations, Environment, CSIRO, Pullenvale, Australia

[2] Science Impact & Policy, CSIRO, Pullenvale, Australia

[3] Robotic Design and Interaction, Data61, CSIRO, Pullenvale, Australia

Published online: 08 August 2024

ethical risk. From the philosophy of technology, we draw on the understanding of AI systems as sociotechnical systems that include the physical artifacts that make up the AI itself, the stakeholders who affect and are affected by it, and the institutions that determine how it is used. We add to this the recognition that ethical responsibilities may take different forms (such as obligation, accountability, and blameworthiness), and that the stakeholders within these systems have ethical responsibilities to other stakeholders who may affect or are affected by using the AI. From the ethics of risk, we draw on the insight that stakeholders may be decision makers about a risk, beneficiaries of it, exposed to risk, or some combination of these roles [7]. Finally, we use the concept of domination from republican political theory to analyse whether the relationships between stakeholders that these risk roles describe create circumstances where a stakeholder is negatively affected by the decisions about an AI system made by another without having a say in that decision.

Drawing on these concepts allows us to elaborate on how the roles played by different type of stakeholders (such as developers, end users, and data subjects) describe their ethical responsibilities towards others, and how they are dependent on the actions performed by other stakeholders. Emphasising the relationships between stakeholders gives this account of ethical risk for AI a clear way of distinguishing ethical risk from other forms of risk. This would assist developers, users, and those affected by AI in recognising the ethical risks of AI applications that do not necessarily correspond to more recognisable legal and technical risks. Stakeholders may find this approach useful to avoid incurring 'ethical debt' in AI, where AI systems are designed, developed, deployed, and used without anticipating the potential ethical issues with the system [8].[1]

To support our account of ethical risk for AI, in Sect. 2 we briefly survey how ethical risk has been defined previously in business and professional ethics, and in technical standards. In Sect. 3 we explain why AI presents a particular problem for how we usually understand risk and responsibility for technology. In Sect. 4 we discuss how AI systems may be understood as sociotechnical systems that include the physical artifacts that make up the AI itself, the stakeholders who affect and are affected by it, and the institutions that determine how it is used. We also arrive at our new definition of ethical risk for AI, and introduce the two examples of AI applications we will use to illustrate our discussion of the relationships between AI stakeholders. In Sects. 5 and 6 we discuss how identifying whether stakeholders are decision-makers about a risk or are affected (positively or negatively) may be used to identify dependent relationships between them. In Sect. 7, we argue that the ethical responsibilities of stakeholders who are decision-makers about how others are affected by the risks posed by an AI system prevent these dependent relationships from becoming ones where the decision-maker dominates those dependent on their decisions. Finally, in Sect. 8 we conclude the paper with a summary of what we have presented.

2 The problem of AI for ethical risk

The ethics of risk is a rich vein of ethics and political philosophy, and due to space constraints we can only give a brief sketch of the concepts that are applicable to AI here.[2] Ethical risk is a concern for AI for several reasons: Blackman [6] lists physical harm, mental harm, autonomy (privacy violations), trustworthiness and respect, relationships and social cohesion, social justice and fairness, and unintended consequences as categories of ethical risk for organisations that use or develop AI. AI technology and its applications have several characteristics that contribute to these risks, such as:

- Many forms of machine learning (ML) are opaque (or 'black boxes'), meaning that it is difficult (if not impossible) for developers and users to understand exactly how the system made a particular decision [9, 10]. This opaqueness may affect the trustworthiness of AI systems.
- AI may be incorporated into robotic systems (such as autonomous vehicles), where the AI is in control of the system's physical functions. This may result in physical harm if the AI makes errors in deciding how to operate the physical systems that implement its decisions.
- Implicit and explicit biases in the data used to train ML systems may be reflected in the system's output [3, 11, 12]. Such biases may negatively affect trustworthiness and respect, social cohesion, and social justice and fairness.
- Using AI systems for decision-making may lead to 'responsibility gaps', where there is uncertainty about responsibility for the decisions recommended (or actions taken) by AI systems [13, 14]. This uncertainty is created by situations where it seems appropriate to hold someone responsible for the output of an AI system, but there are legitimate doubts about who (if anyone) is responsible the system producing that output [15].
- Using AI systems to make decisions about people's lives and livelihoods, without providing them the information

[1] Ethical debt is analogous to the concept of 'technical debt' in software development, where simpler but less robust solutions to technical problems are used with the intent of replacing them with more robust but complex solutions later [8].

[2] Detailed examinations of the ethics of risk include Lewens [72], Hansson [70], and Nihlén Fahlquist [31].

needed for them to advocate for themselves in response [1, 16].

- AI systems may be used to optimise user engagement with a social media platform, leading to platforms showing users more extreme content to keep them engaged. Users may become radicalised by the growing amount of extreme content (that also encourages distrust of alternative views) they are exposed to via the platform, creating 'echo chambers' for these users [17, 18]. This may cause mental harm to users and damage human relationships and social cohesion.

While Blackman does not explicitly define ethical risk, he describes mitigating ethical risks as a means of avoiding breaches of an organisation's ethical values.[3] These values may be clarified by considering 'ethical nightmares' that an organisation should strive to avoid [6]. Such ethical nightmares may be derived from the organisation's context, such as the industry it belongs to, the kind of organisation it is, and its relationships with stakeholders [6].

Defining ethical risk for AI is important for distinguishing this type of risk from other risks that may emerge from developing or using AI. Ethical risk is mentioned in business and professional ethics. To give just two examples, it is described as "simply risks that occur from ethical failures" [19], and "the risk of consciously adopting unethical behaviour" [20]. In this context, ethical failures may be unwanted events that are the result of unethical behaviour, or be the cause of such events, either directly or through omission. While Rotta [20] states that while individual companies will define what ethical risk means to them, it includes compliance risk (failing to abide by relevant laws and regulations and internal policies and procedures), fraud risk, and reputational risk (the effect on the company image from negative events). These accounts describe ethical risk as the possibility of wrongdoing by business employees and executives. It suggests a connection between ethical risk and the responsibilities of those who perform business functions, i.e. their *role responsibilities*.

Guidelines and legislation for appropriate uses of AI are another potential source for defining ethical AI. One common approach is to distinguish between different levels of risk from AI. The EU AI Act, which defines risk as "the combination of the probability of an occurrence of harm and the severity of that harm", contains a list of high-risk AI applications that include biometrics, critical infrastructure, employment, access to essential services, justice administration, and in law enforcement [21].[4] The AI Act also refers to the ethical principles contained within the High-Level Expert Group on AI's Ethics Guidelines for Trustworthy AI [22]. While this may suggest that ethical risk for AI is the risk that its development and use may be contrary to these guidelines, neither the Ethics Guidelines or the Act define 'ethical risk'.

We may also look towards technical standards as a source for a definition. For example, IEEE *Standard 7000: Model Process for Addressing Ethical Concerns during System Design* defines ethical risk as "[a] risk to ethical values", where 'risk' is defined as the "effect of uncertainty on objectives" and 'ethical values' is defined as a "[v]alue in the context of human culture that supports a judgment on what is right or wrong" [23]. In a paper on the ethics of robotics and AI, Winfield and Winkle [24] quote the British Standards Institute (BSI) definition of 'ethical risk' from standard BS8611-2016 *Guide to the Ethical Design and Application of Robots and Robotic Systems* as the "probability of ethical harm occurring from the frequency and severity of exposure to a hazard". 'Ethical harms', defined in the same standard, are "anything likely to compromise psychological and/or societal and environmental well-being" [24]. This definition presents ethical risk as a broad category that includes psychological, societal, and environmental risks.

This definition captures major ethical concerns that surround risks; in particular, how they may cause psychological, societal, or environmental harm. As Blackman's list of categories for ethical risk from AI suggests, mental (or psychological) harms, and damage to social cohesion are within the scope of ethical risk for AI. However, the BSI definition emphasises a consequentialist understanding of ethical risk by focusing on potential harms. Other accounts of ethics (such as deontological ethics) place a stronger emphasis on the significance of ethical wrongs, which are actions that undermine or ignore intrinsic goods (such as respect for persons and their rights), regardless of whether harm has occurred. Harms may not necessarily be wrongs and vice versa [25]. For example, driving a vehicle recklessly but without hitting anyone or causing damage may be a wrong, but not necessarily a harm [26].

We might address the concern that focusing on harms overlooks ethical wrongs by equating harms with wrongs. We may interpret wrongs such as rights infringements as harms to psychological or social well-being, as ignoring rights may be psychological harmful or lead to discrimination or disrespect that is socially harmful. Similarly, we may consider the potential impact of an AI system on the

[3] This does create the possibility for an organisation's values being out of alignment with those of the society in which it operates, especially if they are unconcerned about reputational risk or believe that such risk can be manageable.

[4] A similar approach that distinguishes between low, medium, and high risk applications (based on the potential impact of the application, its duration, and its reversibility) has been proposed by the Australian government [73].

human rights of those affected by its use. The risks of AI to human rights are using it to violate human rights, failing to consider human rights during the AI's design, and the negative impacts on human rights by using AI [27]. The wrongs of violating human rights may be regarded as harms. Similarly, if animals and nature are regarded as having rights, infringing these rights may be understood as harm to environmental well-being. However, Blackman [6] cautions against equating harms with wrongs as this obscures cases where harms and wrongs do not overlap. Similarly, describing the imposition of risk as a harm is not without its difficulties, and does not adequately account for the wrongness of imposing a risk that does not result in harm occurring [26]. For example, using an AI system may be considered an ethical risk if it is used to make decisions or perform tasks that we regard as being human responsibilities. This might be because there is the possibility of moral deskilling (where our abilities of forming moral judgements are negatively affected) if we rely on AI to perform tasks that are our ethical responsibilities [28]. There are also often power imbalances between the stakeholders associated with an AI system, especially if those affected by the decisions made using an AI system have no role in deciding how the system is used.

We may draw several points from these accounts of ethical risk. Ethical risks relate to the possibility of unethical behaviour, and it may also cover social, environmental, and psychological harms. The discussions of ethical risk by Rotta and Blackman suggest that the responsibilities of stakeholders are significant for identifying ethical risks. The BSI definition highlights that ethical risk should cover both ethical harms and ethical wrongs, so that it captures both actual and potential exposure to harm, and unacceptable exposure to risk. As we describe in the next section, we will build on the connection between ethical risk and the responsibilities of stakeholders.

3 Ethical responsibility and ethical risk

Risk and responsibility are intertwined [29–31]. Modern societies see risk as something to be controlled and managed, implying that someone should *take responsibility* for managing and controlling it or be *held responsible* if it occurs [31]. Attributions of responsibility should also be *fair* in that those responsible are aware of what they are doing (or not doing) and are free to act on their responsibility, and *effective* in that they encourage the actions and behaviour they are intended to foster [32].

We have already mentioned responsibility gaps as one of the ethical risks associated with AI applications. To explain why these gaps are a concern, we must first elaborate on the different kinds of responsibility. Ethical or moral responsibility are *normative* forms of responsibility, which may follow from *descriptive* forms of responsibility [33, 34]. Descriptive forms of responsibility (which describe who or what *is* responsible for something occurring) relevant to our discussion are causal responsibility, role responsibility, responsibility-as-authority, and responsibility-as-capacity [33, 35]. People may be casually responsible for an action (causal responsibility), perform roles with designated functions (role responsibility), and may also have authority to take responsibility for the actions of others or for an organisation if they have a position of authority (responsibility-as-authority). Responsibility-as-capacity is another form of descriptive responsibility as it describes whether someone or something possesses *moral agency*, the capacity to perform ethical reasoning and act upon it. Causal responsibility for an action or event may or may not correspond to ethical responsibility: a storm may be causally responsible for damaging a house, but it is not ethically responsible as it does not have moral agency.

Forms of normative responsibility are obligation, accountability, blameworthiness, and liability, as they describe who *should* be held responsible in some form. Obligations are duties to ensure that a stated action or situation occurs in the future [34]. Accountability is the responsibility to explain to others (or give an account for) why a certain action occurred (or did not occur) [34]. Accountability is important for gaining and maintaining trust in technology developers, understanding the causes of technical problems, and how to avoid similar problems in the future [36]. Blameworthiness identifies whether an agent may be morally rebuked for an action [34]. Liability is the duty to compensate those affected by an action or event [34].

Normative forms of responsibility may extend to past events or to possible events now and in the future. Being held accountable, liable, or being considered blameworthy for an unwanted event are backward-looking responsibilities, as they refer to past inabilities to manage or mitigate risk. Having an obligation to prevent and mitigate future risks, and being accountable, blameworthy, or liable for events that may occur are forward-looking responsibilities [35].

Liability may be distinguished into moral liability (a duty to remedy or compensate for an action or inaction) and legal liability (an obligation to be punished or pay damages for an action) [34]. Whether someone should be regarded as morally liable depends on whether they are blameworthy for an action or inaction [34]. As legal liability overlaps with legal risk (understood as exposure to legal claims of damages, compensation, or infringement of laws or regulations) [37], we will not consider it further here.

Normative responsibilities may (but not necessarily) follow from descriptive responsibilities. Causal responsibility, moral agency (responsibility-as-capacity), and the potential for wrongdoing are preconditions for accountability [34]. Having a role responsibility may bring with it obligations and a duty to be accountable for one's actions. The conditions for a reasonable attribution of blameworthiness to an agent are moral agency, causal responsibility, the action was freely performed and with knowledge of its likely effects, and that wrongdoing has occurred [30, 34].

Failing to fulfil an ethical responsibility is an ethical wrong as it is a failure to fulfil an ethical duty. Failing an obligation towards another is a wrong as it is a failure to uphold an ethical duty towards them. Similarly, failing to be accountable, blameworthy, or liable is a wrong as it is a failure to accept an ethical duty. Failing to fulfil a responsibility may also be a harm if that wrong is a setback, thwarting, or denial of someone's interests [25].

How do these concepts of responsibility apply to AI systems? AI may possess descriptive responsibility if it is casually responsible for an action or if it is used in a role where it performs a designated function. Whether AI systems may possess responsibility-as-capacity or moral agency, and if so, whether it is in isolation of the moral agency of the human agents associated with it, is still debated [38, 39]. In this paper we will assume that moral agency (and thus, ethical responsibility) is only possessed by human persons, and that AI systems as artificial agents do not possess moral agency [40]. If AI systems possess only casual responsibility, they do not possess any form of normative responsibility, since normative responsibilities require an agent to possess responsibility-as-capacity or moral agency. The moral agents involved in the decisions and actions made by AI systems would be the human agents involved with the system, such as the users and developers. AI would appear to be just another technology when it comes to ethical responsibility.

Technologies poses two major questions for ethical responsibility: whether technology developers have special responsibilities, and whether using a technology affects the responsibilities of its users [32]. Given the connection between responsibility and risk, AI developers appear to have a *prima facie* responsibility to control and manage the risks connected to the AI system. The legitimacy of attributing to developers this *prima facie* responsibility to control and manage these risks depends on whether it is fair and effective to do so. The fairness of this attribution depends on the developer's capability to be aware that the risk exists and its likely impact, and on their ability to decide whether to accept that risk.

The fairness of attributing responsibility to technology developers is often tied to a 'control requirement': the developer is rightly held responsible for the actions of the system if they have control over it [13, 41]. While AI systems are artifacts created by humans for human purposes (like other technologies), AI systems are *artificial agents* with properties that other technical artifacts lack, such as capabilities for autonomous decision making and interacting with their environment [42]. This agency creates uncertainty over how the system will respond to the inputs it receives. This uncertainty is further compounded as machine learning (ML) models (currently the dominant approach to implementing AI) implement algorithms that they develop themselves based on the training data that they process [43]. While developers can control the training data used by ML models, the developers cannot predict the model's output. As a result, developers have less control over the AI system than they would have over traditional computing systems where developers implement the algorithms within the system themselves. This reduced control developers have over AI compared to other technologies may contribute to responsibility gaps.

Responsibility gaps may be distinguished into four varieties: culpability, moral accountability, public accountability, and active responsibility [14]. Culpability gaps occur where blameworthiness for an AI system's actions or decisions cannot be attributed to its developers or users. Both moral and public accountability gaps refer to the inability of those relying on the recommendations of AI systems to explain how the system arrived at that recommendation. The difference between moral and public accountability gaps are the audience for the explanation: moral accountability is an individual's account of their actions or decisions to other individuals, while public accountability is a public official or role holder's account of their actions or decisions to those affected by them. Moral accountability follows from general ethical responsibilities of stakeholders as moral agents, while public accountability may follow from the ethical responsibilities of a stakeholder's occupational role (we discuss occupational roles further in Sect. 4). Active responsibility gaps occur when developers and users of AI systems are unaware of their obligations to those who may be affected by the system, and where developers and users may be unable or insufficiently motivated to fulfil these obligations [14].

As this distinction between different varieties of responsibility gaps suggests, the lack of control developers have over AI systems does not necessarily mean that they do not have some form of ethical responsibility for these systems. Developers still have control over how they mitigate the potential risks of the system and whether they follow relevant regulations and guidelines in developing the AI [41]. We may also consider developers to be accountable for the actions of their systems, and to have an obligation

to account for their system's actions in the future [15, 44]. Being an AI developer carries with it a role responsibility to be accountable for how their system performs, and an obligation to provide such explanations to other stakeholders in the future if needed. As we are considering AI developers as stakeholders rather than individual persons, this is a form of public accountability to other stakeholders.

Whether attributing responsibility to developers is effective depends on the behaviours it is intended to promote. Ideally, attributing ethical responsibility to developers for the risks associated with the technology they create will encourage them to address these risks, through mitigation, management, or removal. In terms of the varieties of responsibility gap described above, attributing ethical responsibility to AI developers should prevent active responsibility gaps caused by being unaware of their obligations towards those affected by the outputs created by their AI systems. AI developers are also not the only stakeholders who may be affected by active responsibility gaps: the users of AI systems may also be unaware of the responsibilities they have towards other stakeholders.

Based on this discussion of AI, ethical risk, and ethical responsibility, we define *ethical risk for AI* as the possibility that a stakeholder connected to that AI system may fail to fulfil one or more of their ethical responsibilities towards another stakeholder. By defining ethical risk as a failure by a stakeholder to fulfil an ethical responsibility, an ethical risk is a potential ethical wrong that may also be an ethical harm if the risk occurs. Similarly, a stakeholder is *at ethical risk* from an AI system if they are dependent on a stakeholder who makes decisions about the AI,[5] and so may be wronged or harmed by the decision-maker's failure to fulfil their ethical responsibilities towards them.

To elaborate on this account, we will draw on the conception of AI systems as sociotechnical systems to describe how stakeholders relate to the technical artifacts and technical norms that compromise an AI system, and then elaborate on how stakeholders have different roles in their relationship with the AI. We then discuss how these roles may create dependency relationships between stakeholders. A stakeholder's dependency on another stakeholder's decisions about a risk places them at ethical risk. We will then discuss how the decision-making stakeholder's ethical responsibilities serve to prevent a dependency relationship from becoming one of domination by the decision-maker. Throughout the rest of this paper, we will use two examples of AI systems that pose ethical risks to illustrate our discussion:

- *Bespoke surgical tools*: A generative AI system that designs attachments for surgical robots that are optimised for use on a specific patient by a surgeon for a specific operation [45].
- *Robotaxis*: A company operates self-driving cars as a taxi service within a suburban environment, where the vehicle's passengers are not expected to intervene in its operation [5].

4 AI as sociotechnical systems

The role of stakeholders in an AI system is best understood by recognising AI as a *sociotechnical system* [46]. Sociotechnical systems are hybrid systems that include both physical artifacts and human elements such as individuals, organisations, and institutions that affect how these artifacts are used [47]. Sociotechnical systems traditionally contain three types of components: technical artifacts, human agents, and institutions [42, 47]. AI systems also include two additional components: artificial agents and technical norms [42]. Technical norms are rules that are implemented within AI systems, either by the developer or developed by AI systems themselves from analysing training data and/or from the environment they operate in [42]. AI developers have ethical responsibilities (within the constraints of fairness and effectiveness stated above) for the software and hardware elements of the AI system (i.e. the technical artifact of the AI), the artificial agent (the AI when it is operating) and the rules encoded into the AI (i.e. the technical norms) that affects its decisions and actions.

The technical artifact is the hardware and software that comprises an AI system. Recognising the hardware necessary to develop and use an AI system ensures that the material characteristics of AI (such as the environmental impact and financial cost of operating it) are not overlooked [48]. The software includes the AI model performing the classification, prediction, or decision-making, and the other software necessary for it to operate (such as the operating system that the AI model runs on). When the AI model is operating, it may be considered as an artificial agent. For bespoke surgical tools, the technical artifact is the hardware and software used to operate the generative AI that uses patient scans to design an optimised shape for an attachment for a surgical robot to perform an operation on that patient. The combination of the software that performs the autonomous driving of the vehicle and the vehicle itself is the technical artifact for the robotaxi.

The human agents, or stakeholders, are anyone who may affect and is affected by an AI system, either directly or indirectly [49, 50]. Stakeholders may be identified by their role

[5] As we will explain in Sect. 4 on AI as sociotechnical systems, these are decisions about the *technical artifacts* (the hardware and software elements of the AI) and the *technical norms* (the rules encoded into the AI) of that system.

in interacting with (or being affected by) an AI system [50]. The stakeholder's role may describe their duties in relation to an AI system, their contextual identity, or to the circumstances in which that system affects them [50]. The role of developer, for instance, designates the individuals or groups who create and design a particular AI system. Similarly, the workers who prepare the data used to train AI systems [51] are also stakeholders. A user is the contextual identity of someone intentionally using an AI system for a given purpose. Cyclists and pedestrians are stakeholders in self-driving vehicle technology, as they are likely to be present and affected by its use.

Mapping the process that the AI will be used in is one method of determining the stakeholders who interact with the AI directly, those who are indirectly affected by the use of AI by others, and those who supply data that the AI uses [52]. For example, a case study of AI-designed attachments for surgical tools identified eight stakeholders who may affect or are affected by the AI system that designs these tools [52]:

- designers who develop the AI system;
- fabricators who use 3D printing to create the AI-designed tool;
- hospitals and medical institutions where the operation using these tools takes place;
- patients who are treated using the bespoke surgical tool;
- radiologists who perform the patient scans that the AI uses as input to create a bespoke surgical tool design for that patient's operation;
- regulators who determine what tools, technologies, and techniques are permissible to use in healthcare settings;
- surgeons who decide to use a bespoke surgical tool to treat their patients; and.
- surgical colleges who control the certification of surgeons and determine what tools, technologies, and techniques they are permitted to use.

Of these eight, fabricators, patients, radiologists, and surgeons directly affect the AI or are directly affected by its use. Surgeons decide whether to use it to create a specialised tool for treating an individual patient. Radiologists provide the patient scans that the surgeon uses as input for the AI system, and the output of the AI system is the bespoke surgical tool design is the design the fabricator uses to create the tool using 3D printing. Patients are treated using a tool it has designed.

The remaining stakeholders (designers, hospitals and medical institutions, regulators, and surgical colleges) indirectly affect the use of the AI. Hospitals and medical institutions, regulators, and surgical colleges indirectly affect the AI by imposing rules, regulations, and policies that determine how it is used. Designers are indirectly affected by other stakeholders using the AI they have developed as their commercial success, reputation, and legal liability will be impacted by the quality of the surgical tools designed by the AI they have created.

The different types of stakeholders may also be distinguished into classes depending on their relationship towards it. The literature on the stakeholders for interpretable and explainable AI is a good starting point for this purpose [53, 54]. While this is not intended to be an exhaustive list, possible classes of AI stakeholders include:

- *Accident and Incident Investigators*: those who investigate failures and accidents involving AI systems to determine whether the AI system was casually responsible [54]. Hospitals and medical institutions, regulators, and surgical colleges who investigate potential failures of surgical tools designed using AI, and the investigators of vehicle incidents involving robotaxis belong to this class of stakeholders.
- *Data-Preparers*: those who generate and annotate data used as the training data for developing ML models [51].
- *Data-Subjects*: those whose personal data is contained in the training data used to train a ML model [53].
- *Developers/Service Providers*: those who develop and support an AI system [54]. They also be distinguished between owners of the AI system's intellectual property, and the implementers who develop the system itself [53].
- *End-Users*: direct users of an AI system who are directly affected by it [54]. Surgeons are end-users of AI-designed surgical tools as they directly use these tools to treat their patients. The passengers who use robotaxis are also end-users.
- *Expert Users*: direct users of an AI system who are indirectly affected by it [54]. The radiologists who provide patient scans for the AI are expert users, as they are only indirectly affected by how well the AI designed the bespoke surgical tool.
- *Insurers*: those who cover financial risks for developers and operators of AI systems [54]. Medical insurers would decide whether they are willing to accept covering financial costs for the potential risks of surgeons using bespoke surgical tools. Vehicle insurers would also consider whether they are willing to accept the financial costs of liability claims for accidents caused by robotaxis.
- *Prediction-Recipients*: those directly affected by the decisions and predictions made by an AI system, but are not users of the system themselves [54]. Patients treated using an AI-designed surgical tool and the fabricators who use 3D printing to create that tool are both

prediction-recipients. For robotaxis, other road users and pedestrians are prediction-recipients.

- *Regulatory Agencies*: those who protect the interests of those directly affected by an AI system (such as prediction-recipients and end-users) [54]. Belonging to this class are the regulators and surgical colleges who determine whether an AI system for designing surgical tools may be used, and the regulators who determine whether self-driving vehicles are permitted on public roads.

These possible classes offer a starting point for identifying the stakeholders who relate to a specific AI system.[6] These stakeholder classes also have interactions between themselves: the decisions made by developers, for example, will affect end-users, expert users, and prediction-recipients.

Institutions are rules, laws, social norms, and regulations that stakeholders follow in their actions and decisions [47]. For bespoke surgical tools, institutions include the legal and professional requirements for surgical operations, the professional ethics of surgeons and medical staff, best practice guidelines for surgery, and the regulations and procedures of hospitals and medical institutions. For robotaxis, institutions include the laws and regulations that govern both motor vehicle use and the use of autonomous vehicles on public roads, and the safety requirements for vehicles. These rules and regulations may also define the roles of stakeholders [47]. We can distinguish between the general and role responsibilities of stakeholders [55].[7] General ethical responsibilities are ethical duties held by any agent that possesses moral agency. All stakeholders share the same general ethical responsibilities that accompany moral agency. Role responsibilities are duties that follow from being a particular type of stakeholder, such as a doctor, electrical engineer, or software developer. These duties may be part of the stakeholder's *occupational role*, which is a form of social role where the role holder internalises a set of attitudes associated with that role and acts in ways expected of those who perform it [56]. Professions and professional organisations are institutions that define the occupational role of their members. Professions may define these duties in codes of conduct or standards that members of an occupational role must adhere to as part of their professional accreditation. Professional organisations that set standards and expectations of their members may also be stakeholders if they have a say in whether (and how) their members use AI. The IEEE standard that presents a description of ethical risk that we mentioned earlier (Standard 7000 *Model Process for Addressing Ethical Concerns during System Design*) is one example of how professional organisations may play a role in how AI is designed and used [23].

Other stakeholders, such as users and patients, are social roles rather than occupational roles. Their social role is largely defined by their interactions with an AI system or with other stakeholders who are affected by its use. Consider some of the stakeholders for an AI system that designs bespoke surgical tools. The surgeons and radiologists are occupational role stakeholders, while the patients are social role stakeholders, as 'patient' is the contextual identity of persons seeking and undergoing medical treatment. Social roles are not necessarily distinct from occupational roles: an injured soldier has both the occupational role of soldier and social role of patient.

As mentioned above, the artificial agent is the AI itself when it is operating as part of the technical artifact, and although it has agency (as it can make decisions and possibly interact with its environment if it is programmed to do so), we have assumed for this paper that it does not possess moral agency (or responsibility-as-capacity). As such, only the stakeholders that are part of the AI sociotechnical system possess ethical responsibility. However, the artificial agent may possess *causal* responsibility as it is the direct cause of the actions or decisions performed by the AI system. For bespoke surgical tools, the artificial agent is the software that uses a patient scan to determine an optimal design for an attachment for a surgical robot to perform a specific operation on that patient. For robotaxis, the artificial agent is the autonomous driving software that is in control of the vehicle that carries passengers to their destination.

Depending on how it is implemented, the technical norms within an AI system (i.e. the rules that determine how it makes decisions) are defined by the developer, determined by the AI itself through analysing training data or by analysing or interacting with its environment [42]. AI systems that are implemented using symbolic AI (so-called 'good old-fashioned AI' or GOFAI) use formal models of the AI's operating environment and heuristics to make decisions [57]. In such systems, the technical norms are defined by the developer, and the developer can understand (in principle) how the system made a specific decision. However, the most effective AI systems in real-world applications are ML systems, where the algorithms that determine how the AI makes decisions are developed by the AI itself through processing training data or some other method such as evolutionary algorithms [43, 58]. In these cases, the technical

[6] Different stakeholder classes may also have different perceptions of risk, and may identify risks relevant to them that others may overlook or disregard. Including multiple perspectives of risk is an important method of avoiding 'professional ethnocentrism' of engineers and other technical fields that priorities objective measures of risk over public risk perceptions [74]. We will not explore this point further here.

[7] Alexandra and Miller [56] make a similar distinction between internal and external responsibilities, where internal responsibilities correspond with role responsibilities, and external responsibilities with general responsibilities.

norms are determined by the artificial agent rather than the developer directly, and the developer may not be able to explain why the AI system made a specific decision. However, the developer still has some control over the possible decisions the AI may make by imposing restrictions on the permissible decisions it can make. For example, the AI that designs bespoke surgical tools may have restrictions on the dimensions and characteristics that the tools it designs may have. Similarly, a robotaxi may have technical norms that prevent it from making certain kinds of decisions, such as ignoring traffic signs.

5 Stakeholders and risk roles

The classes of AI stakeholders mentioned in Sect. 4 may be used as a starting point for identifying the stakeholders involved with a particular AI system. An AI system will necessarily have developers and service providers. Those who use the AI system may be end-users if they are directly affected by their use of the system or expert users if they are only indirectly affected by their use of it. For example, someone using an AI to recommend films for them to watch is an end-user, while a professional using an AI to assist them in making decisions for a client is an expert user. The client of an expert user is a prediction-recipient, as they are affected by the output of the AI but do not use it themselves. Depending on what the AI is being used for, there may also be regulators who control how the AI may be used, as well as incident investigators who determine whether the AI was casually responsible for harmful outcomes.

There are considerable interactions between these stakeholders. These interactions may include ethical responsibilities of obligation, accountability, blameworthiness, and liability. While the specific ethical responsibilities will depend on the individual AI system, there are some general responsibilities that are likely to exist between the stakeholder classes associated with an AI. End-users may be accountable for how they use AI. Expert users will have obligations towards and be accountable to prediction-recipients. Developers and service providers will be accountable to regulators and incident investigators. Incident investigators will establish whether any stakeholders are blameworthy for a harmful or wrongful use of AI, and regulators will determine whether any stakeholders are liable for such use.

The specific ethical responsibilities between stakeholders may be identified by considering how an AI system is used within a larger process. How stakeholders fulfil their responsibilities will affect other stakeholders and their ability to fulfil their own responsibilities in the process. Mapping out the process in which AI is used will identify how stakeholders make decisions about using it, provide resources for it to operate, utilise the system for a given purpose, and evaluate its effectiveness [52].

The relationships between stakeholders may be further clarified by identifying the risk roles each stakeholder holds. Any risk has associated with it the roles of *beneficiary*, *decision-maker*, and *risk-exposed* [7, 59]. Beneficiaries gain from the risks taken either by themselves (which also gives them the roles of decision-maker and risk-exposed for that risk) or by others. For risks where there is no benefit to those affected and where the costs of prevention, mitigation, and recovery are not borne by the risk-exposed themselves, the 'beneficiaries' are those who bear these costs [7].[8]

The risk role of decision-maker corresponds with being ethically responsible in some form (such as having an obligation or being accountable, blameworthy, or liable) for that risk. As they possess ethical responsibility, the decision-maker will necessarily possess responsibility-as-capacity (i.e. moral agency). The ethical responsibilities of being the decision-maker may be for risk reduction, risk assessment, risk management, or risk communication [60]. Each of these responsibilities are obligations as they refer to reducing, assessing, managing, or communicating risks that may occur now or in the future. These responsibilities may be part of a broader framework of risk governance [61, 62].

The relationship between the risk role holders (beneficiary, decision-maker, and risk-exposed) describes the responsibilities that exist between them. After a risk has occurred, the decision-maker may be:

- *accountable* to those who may have benefited from it or were exposed to that risk for why they decided to take it,
- *blameworthy* if deciding to take that risk was not ethically permissible (either for themselves or to the beneficiaries and the risk-exposed), or.
- *liable* if they should be punished or compensate the risk-exposed or potential beneficiaries for deciding to take that risk.

For example, a surgeon using an AI-designed tool to treat a patient is accountable to that patient for this decision, and the developer of the generative AI system that designs these tools is accountable to surgeons, patients, the fabricators who use 3D printing to create these tools, and the radiologists who perform the patient scans used as input for the AI [63]. Similarly, the robotaxi developer is accountable to the passengers of their vehicles, and to other road users for how their vehicle operates. In both cases, the AI developer is accountable to the relevant regulators.

[8] Hansson [7] suggests the term 'counter-affected' in place of 'beneficiary' to better describe this role. For simplicity, we will use 'beneficiary' to describe both beneficiaries and the counter-affected.

To distinguish between accountability and blameworthiness, we must consider the conditions for accountability and potential reasons for why blameworthiness does not follow from being accountable [35]. As mentioned earlier, the conditions for accountability include being a moral agent (i.e., responsibility-as-capacity), being the cause of an event (i.e., being causally responsible), and the potential for wrongdoing to have occurred [34]. If these conditions are met, the accountable agent may not be blameworthy if they lack the knowledge (or could reasonably have had the knowledge) of the outcome of their action (or inaction), if they were not free to choose their actions (or inactions), and if no wrongdoing has occurred. As blameworthiness is a precondition for moral liability, a blameworthy decision-maker may also be morally liable: it would be appropriate morally for them to compensate the beneficiary or risk-exposed (or both) for their management of the risk.

In the bespoke surgical tool example, a surgeon may be blameworthy if they did not attempt to mitigate the risks of using an AI-designed tool in surgery. A surgeon may mitigate these risks in several ways. They may consult with the radiologist performing the patient scans used as input for the AI system that designs the tool that the scans are correct, and they may consult with the operator of the 3D printer that creates the tool that it is fit-for-purpose before it is used in surgery, and inspect the tool themselves before they use it [63]. The developer of the generative AI system that designs the tool may also be blameworthy if they do not the mitigate the risks of their system designing a tool that is unfit for use by the surgeon [63]. The blameworthiness (and potentially, the liability) of these stakeholders would be determined by stakeholders who are accident and incident investigators, such as hospitals and medical institutions, regulators, and professional organisations.

In the robotaxi case, if a robotaxi hit a pedestrian, the vehicle's developer would be accountable to the other stakeholders to explain why the risk occurred.[9] For the developer to also be blameworthy, it would also have to be established that they could have reasonably foreseen the circumstances in which the accident occurred, and that there were effective means of mitigating this risk that were not used or had been ineffective. Establishing this is the role of the accident and incident investigator. The developer may also be morally liable if it would be legitimate to punish them for failing to prevent the risk from occurring or if they have a duty to compensate those affected by the accident.

Identifying the relationships between risk role holders highlights where these risk roles overlap, and where the relationship between these roles is a dependence relationship [7]. Risk roles overlap if one party holds two or all three of the roles of beneficiary, decision-maker, and risk-exposed. A dependence relationship occurs where one risk role is dependent on another for something of value. This dependence (where one party can exercise power over another) is ethically significant as it may indicate that the dependent party lacks autonomy over their exposure to risk, or that they may be exploited for another's benefit.

6 Risk relationships and dependency

A variety of relationships may exist between the stakeholders affected by an AI's ethical risks. Combined with identifying the roles stakeholders possess in relation to an AI, clarifying the type of relationships that exist between stakeholders will indicate the types of ethical responsibilities they have to each other. Wolff [64] describes five relationship types that may exist between decision-makers, the beneficiaries of taking a risk, and the risk-exposed: individualism, paternalism, maternalism, externality, and adjudicatory. Individualism is the simplest relationship, as there is only one party who is the beneficiary, decision-maker, and exposed to risk (i.e. there is complete overlap in who holds all the risk roles) [64]. As we are concerned with ethical responsibilities to other stakeholders, we will not consider this type of relationship further here.

The other relationships (where the risk roles of beneficiary, decision-maker, and risk-exposed are distributed across two or more stakeholders) may potentially be dependent relationships between the decision-maker and the beneficiary or risk-exposed (or both). These relationships may be ethically relevant as a power imbalance between stakeholders may exist where the risk-exposed and/or beneficiary do not have reciprocal power over the risk decision-maker (Maheshwari & Nyholm, 2022). This lack of reciprocity means that the risk-exposed and/or beneficiary must trust the decision-maker not to exploit their vulnerability to the risks under the decision-maker's control. These relationships are listed in Table 1 and are described further below.

A paternalistic relationship may exist if one party is both the beneficiary and the risk-exposed, while another is the decision-maker [64]. This may occur if an AI developer has full control over the risks of their system, and the user has no choice but to accept the AI system as it is. In this case, the user is at ethical risk from the AI developer failing to fulfill their responsibilities to them. For example, with the AI used to design bespoke surgical tools, the AI developer holds the role of decision-maker, and the patient and surgeon are both beneficiaries of using AI and exposed to the risk of using it to design the tool [63]. In this case, the developer has a

[9] This occurred in October 2023 in San Francisco, where a General Motors (GM) Cruise robotaxi hit a pedestrian that had first been hit by a human-operated vehicle [5, 71]. For reasons of space and scope, we will not examine this specific incident here.

Table 1 Dependency relationships (based on Wolff 2010)

Relationship	Stakeholder A	Stakeholder B	Stakeholder C	Dependency
Paternalist	Decision-Maker	Beneficiary and Risk-Exposed		B depends on A
Guarantor ('Maternalist')	Decision-Maker and Risk-Exposed	Beneficiary		B depends on A
Negative Externality	Decision-Maker and Beneficiary	Risk-Exposed		B depends on A
Adjudicatory	Decision-Maker	Beneficiary	Risk-Exposed	B and C depend on A

paternalistic relationship with both surgeons and patients. Similarly, the robotaxi developer has a paternalistic relationship to the vehicle's passengers.

The inverse of a paternalistic relationship exists where one party is the decision-maker and the risk-exposed, while another is the beneficiary. This situation occurs when a stakeholder acts a guarantor for a transaction by another stakeholder [64].[10] A specific example of this is the commitment made by Microsoft to users of its AI tools Github Copilot and Azure OpenAI Service that it will pay any legal costs that users incur if their creations made using these tools are found to infringe copyright [65]. In this case, Microsoft (as the developer) is the decision-maker and risk-exposed, and the end-users and expert-users of these systems are beneficiaries.

Externalities are created by relationships where one party is risk-exposed while another is the decision-maker and the beneficiary [64]. Externalities are the effects economic transactions have on those who are not involved in that transaction and may be positive or negative depending on whether the effects are desirable or not [66]. Given that risks are potentially unwanted events, this relationship only represents negative externalities. Positive externalities may be better represented by a guarantor relationship. Negative externalities raise ethical concerns as the decision-maker gains the potential benefits of risk-taking without also exposing themselves to that risk [64]. This creates a 'moral hazard', where the lack of risk exposure affects the decision-maker's willingness to take risks that they benefit from [64]. In the bespoke surgical tools example, the AI developer is the decision-maker and beneficiary of the use of this system, while the fabricator who uses 3D printing to produce the designed tool is exposed to the risk of the AI producing a flawed design [63]. The relationship between the developer and the fabricator in this case is a negative externality.

[10] While Wolff [64] calls this a maternalistic relationship to emphasise that it is the inverse of paternal relationship, we will call this a guarantor relationship to avoid unintentional implications that there are gendered aspects to these relationships.

An adjudicatory relationship is created where one party is the decision-maker, another is the beneficiary, and another is exposed to the risk [64]. The decision-maker determines how benefits and risks are distributed, without being a beneficiary or exposed to the risks themselves. Regulators may have this relationship with other stakeholders, as their decisions will determine whether other stakeholders will be able to benefit from the risks of a regulated AI, and which stakeholders are exposed to these risks.

7 Risk relationships and ethical responsibilities

Dependency relationships have the potential to foster the domination of the dependent by whoever has power over them [67]. Domination is an important concept in republican political theory, which interprets domination as arbitrary power over others, and freedom as non-domination [68]. Lovett [67] defines a social power as being arbitrary "to the extent that its potential exercise is not externally constrained by effective rules, procedures, or goals that are common knowledge to all persons or groups concerned".

Maheshwari and Nyholm [69] draw on this to define the concept of *dominating risk impositions*, which are relationships between decision-makers who impose risks onto others and those affected by these risks. A dominating risk imposition exists where there is a dependency between the decision-maker and the risk-exposed, there is a power difference between the decision-maker and the risk-exposed, and the decision-maker's ability to impose risk is arbitrary, so that the risk-exposed cannot limit or control the decision-maker's ability to expose them to risk [69]. This power difference is non-arbitrary if the decision-maker's ability to impose risk is limited by effective rules or procedures that both the decision-maker and risk-exposed are aware of, or if the risk-affected themselves have instructed the decision-maker to make the decision about risk on their behalf, and the decision-maker is accountable to the risk-exposed for this decision [68]. The risk-exposed may be wronged in such relationships if their risk exposure further strengthens the dominating relationship the decision-maker has with them or if it creates a new domination relationship where the decision-maker previously did not dominate that aspect of the risk-exposed's life [69].

Maheshwari and Nyholm [69] use trials of self-driving cars as a taxi service in suburban areas as an example of a dominating risk imposition. The passengers in robotaxis are dependent of the decision-makers who developed the AI controlling the vehicle (for simplicity we will assume the AI developer and the car's operator are the same). This dependency relationship between passenger and AI developer is

not a dominating one as the passenger has chosen to ride in the robotaxi, and the AI developer is accountable to the passenger. Operating robotaxis in a suburban area also affects the risks pedestrians and drivers face in using the roads in that area. Other road users are exposed to the risk imposed by the robotaxi's developer to operate their cars in their area, and as the AI developer is the decision-maker about how the robotaxi operates, the other road users are dependent on them. Unlike the robotaxi's passengers, however, the AI developer is in a dominating risk relationship with other road users, as the other road users cannot prevent the developer from operating the robotaxi on their roads (outside of advocating for this to be made illegal), and the developer's decisions in how the robotaxi operates would be arbitrary to the other users outside of the limits imposed by traffic laws.

As the above example suggests, laws may restrict the arbitrary powers that decision-makers have over those who are dependent on them, and so prevent dependency relationships from becoming dominating ones. The responsibilities of decision-makers may also set limits on how they may use the power they possess over others. The commonly known rules, procedures, or goals mentioned in Lovett's definition of arbitrary social power may be part of the decision-maker's role responsibilities and obligations. These ethical responsibilities serve to prevent the dependent relationships that exist between decision-makers and those affected by their decisions from becoming relationships of domination by the decision-maker. The bespoke surgical tools example demonstrates how dependent relationships may be prevented from becoming dominating ones. As noted in Sect. 6, the developer of the AI system that designs surgical tools has a paternalistic relationship with the surgeons who use these tools and the patients treated with them, as the developer makes the decisions about how the system is implemented and the types of possible tools it can design. Surgeons and patients are dependent on the AI developer. To prevent this dependency relationship from being a dominating one, the AI developer will be accountable for how well the tool designs created by the AI fulfill their intended function, and have an obligation to mitigate the risks of using generative AI for this purpose [52]. The developer would also be blameworthy if they fail to fulfill these responsibilities towards patients and surgeons [52]. Forward-looking responsibilities, such as obligations, may be described as a relationship where one party owes it to another to ensure that some action occurs [35]. In the context of the risk relationships described here, the first party is the decision-maker, the other party is the beneficiary or risk-exposed (or both), and the action is managing (through avoiding, reducing, or mitigating) a risk associated with a technology that the first party controls. This relationship represents an obligation between the decision-maker and the other stakeholders. Where the beneficiary or the risk-exposed (or both) of a risk are dependent on a decision-maker to manage it, the decision-maker has an obligation toward them to do so. Failing to meet that obligation (intentionally or otherwise) means that the decision-maker dominates the beneficiary or risk-exposed, as the obligation has failed as a means of limiting the decision-maker's power over them. Neglecting to effectively manage a risk creates an ethical risk for the decision-maker, and this possibility of the decision-maker neglecting their obligation towards them is an ethical risk for the beneficiary or risk-exposed.

Similarly, backward-looking responsibilities, such as accountability and blameworthiness, may also be described as a relationship between the decision-maker and another party who is the beneficiary or risk-exposed or both where it is appropriate for the other party to hold the decision-maker responsible for some managing a risk under the decision-maker's control [35]. The decision-maker is accountable towards (and potentially held blameworthy by) the beneficiary or risk-exposed (or both) for their management of the risk. In the context of robotaxis, the stakeholders are the car's developers, passengers, other road users, and residents.[11] The passengers are the robotaxi's end users. The road users and residents are prediction-recipients who are exposed to the risks of the robotaxi, such as pedestrian recognition. They are also dependent on the robotaxi's developer as the developer makes decisions about these risks. To prevent this dependency from becoming domination, the robotaxi's developers have ethical responsibilities towards the other stakeholders. These ethical responsibilities will be the forward-looking responsibility of obligation (to reduce, mitigate, or remove the risks), and the backward-looking responsibilities of accountability and blameworthiness. The risks of the self-driving car are *ethical risks* for the robotaxi's developer, and the end-users (passengers) and prediction-recipients (other road users and residents) are *at ethical risk* from these risks.

8 Conclusion

Discussions of AI ethics often use the term 'ethical risk' without defining exactly what this term means and what distinguishes it from other forms of risk. In this paper, we have presented a new definition of ethical risk for AI that emphasises the relationship between risk and responsibility occurring within a sociotechnical system comprised of technical artifacts, stakeholders, institutions, artificial agents,

[11] This list is incomplete: for instance, the regulator who establishes the legality of operating self-driving cars on public roads is another stakeholder, for instance. We are focusing on these stakeholders merely for illustrative purposes.

and technical norms. We define the ethical risks of AI as the possibility that a stakeholder connected to that AI system may fail to fulfil one or more of their ethical responsibilities to other stakeholders.

Defining ethical risk for AI in this way highlights the connections that various stakeholders have to an AI system, and how their decisions may affect others. Identifying the roles of stakeholders and their connections to both the AI itself and to other stakeholders provides a clearer view of their responsibilities and exposure to risk. It is useful for developers as it provides them a better understanding of who may be affected by the AI, and who may have an impact in how the AI is used. AI users are also better informed about how they and other stakeholders may affect the AI, and who is responsible for mitigating the various risks associated with it. Similarly, those affected by how others use AI are better positioned to understand who is making decisions regarding how it is designed and used. Recognising the ethical risks of AI before it is deployed and used can prevent foreseeable harms and wrongs from occurring, which benefits all the stakeholders in AI.

Acknowledgements We would like to thank the attendees of the 2023 Forum on Philosophy, Engineering, & Technology (fPET 2023) held in Delft, the Netherlands, and the reviewers for their comments and feedback on earlier versions of this paper. This research was funded by CSIRO's Responsible Innovation Future Science Platform.

Author contributions D.M.D. and J.L. contributed equally to developing the topic and argument of the paper. D.M.D. wrote and revised the text, with contributions by J.L. and D.H.

Funding This research was funded by CSIRO's Responsible Innovation Future Science Platform.
Open access funding provided by CSIRO Library Services.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. Picador, New York (2019)
2. Benjamin, R.: Race after Technology. Polity, Cambridge (2019)
3. Gebru, T.: Race and gender. in In: The Oxford Handbook of Ethics of AI, pp. 253–269. Oxford University Press, New York (2020)
4. DeArman, A.: The Wild, Wild West: A case study of self-driving vehicle testing in Arizona. Arizona Law Rev. **61**, 983–1012 (2019)
5. Koopman, P.: Anatomy of a Robotaxi Crash: Lessons from the Cruise Pedestrian Dragging Mishap, 8 February 2024. [Online]. Available: https://arxiv.org/abs/2402.06046. [Accessed 17 April 2024]
6. Blackman, R.: Ethical Machines. Harvard Business Revie w, (2022)
7. Hansson, S.O.: Ethical risk analysis. In: Hansson, S.O. (ed.) in The Ethics of Technology: Methods and Approaches, pp. 157–171. Rowman & Littlefield, London (2017)
8. Petrozzino, C.: Who pays for ethical debt in AI? AI Ethics. **1**(3), 205–208 (2021)
9. Zednik, C.: Solving the Black Box Problem: A normative Framework for Explainable Artificial Intelligence. Philos. Technol. **34**(2), 265–288 (2021)
10. Burrell, J.: How the machine 'Thinks': Understanding opacity in machine learning algorithms. Big Data Soc. **3**(1), 2053951715622512 (2016)
11. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surveys. **54**(6), 115 (July 2021)
12. Sriniyasan, R., Chander, A.: Biases in AI systems. Commun. ACM. **64**(8), 44–49 (2021)
13. Matthias, A.: The responsibility gap: Ascribing responsibility for the actions of learning Automata. Ethics Inf. Technol. **6**(3), 175–183 (2004)
14. Santoni de Sio, F., Mecacci, G.: Four responsibility gaps with Artificial Intelligence: Why they matter and how to address them. Philos. Technol. **34**(4), 1057–1084 (2021)
15. Tigard, D.W.: There is no techno-responsibility gap. Philos. Technol. **34**(3), 589–607 (2021)
16. Vredenburgh, K.: The right to explanation. J. Political Philos. **30**(2), 209–229 (2022)
17. Nguyen, C.T.: Echo chambers and Epistemic Bubbles. Episteme. **17**(2), 141–161 (2020)
18. Coeckelbergh, M.: The Political Philosophy of AI. Polity, Cambridge (2022)
19. Tricker, B., Tricker, G.: Business Ethics: A Stakeholder, Governance and Risk Approach. Routledge (2014)
20. Rotta, C.P.: A Short Guide to Ethical Risk. Routledge, London and New York (2017)
21. Union, E.: Regulation (EU) No. 2024/1689 of 13 June 2024 (Artificial Intelligence Act)
22. High-Level Expert Group on AI (AI HLEG): Ethics Guidelines for Trustworthy AI. European Commission, Brussels (2019)
23. Society, I.E.E.E.C.: IEEE Standard Model process for addressing ethical concerns during System Design. IEEE, (2021)
24. Winfield, A.F.T., Winkle, K.: RoboTed: A Case Study in Ethical Risk Assessment, in *ICRES 2020: 5th International Conference on Robot Ethics and Standards*, Taipei, Taiwan, (2020)
25. Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics, 7th edn. Oxford University Press, Oxford (2013)
26. Maheshwari, K.: On the harm of imposing risk of harm. Ethical Theory Moral. Pract. pp. 965–980, (2021)

27. Compact, U.N.G.: Artificial Intelligence and Human Rights: Recommendations for Companies, 2024. [Online]. Available: https://unglobalcompact.org/library/6206. [Accessed 2 August 2024]
28. Vallor, S.: Moral Deskilling and Upskilling in a New Machine Age: Reflections on the ambiguous future of Character. Philos. Technol. **28**(1), 107–124 (2015)
29. Giddens, A.: Risk and Responsibility, *The Modern Law Review*, vol. 62, no. 1, pp. 1–10, January (1999)
30. Kermisch, C.: Risk and responsibility: A complex and evolving relationship. Sci Eng. Ethics. **18**(1), 91–102 (2012)
31. Nihlén Fahlquist, J. Moral Responsibility and Risk in Society. Routledge, London and New York (2019)
32. Nihlén Fahlquist, J: Responsibility analysis. In: Hansson, S.O. (ed.) in The Ethics of Technology: Methods and Approaches, pp. 129–142. Rowman & Littlefield, London (2017)
33. Hart, H.L.A.: Punishment and Responsibility: Essays in the Philosophy of Law, 2nd edn. Oxford University Press, Oxford (2008)
34. van de Poel, I.: Moral responsibility. in In: Moral responsibility and the Problem of Many Hands, pp. 12–49. Routledge, New York and London (2015)
35. van de Poel, I.: The Relation Between Forward-Looking and Backward-Looking Responsibility, in *Moral Responsibility*, N. A. Vincent, I. van de Poel and J. van den Hoven, Eds., Dordrecht, Springer, pp. 37–52. (2011)
36. van de Poel, I., Sand, M.: Varieties of Responsibility: Two Problems of Responsible Innovation, *Synthese*, vol. 198, no. Supplment 19, pp. S4769-S4787, (2021)
37. Mahler, T.: Defining Legal Risk, in *Proceedings of the Conference Commercial Contracting for Strategic Advantage - Potentials and Prospects*, Turku, Finland, (2007)
38. Coeckelbergh, M.: AI Ethics., A.I.: Cambridge, Massachusetts: The MIT Press, (2020)
39. Gunkel, D.J.: Mind the gap: Responsible Robotics and the problem of responsibility. Ethics Inf. Technol., (2017)
40. Johnson, D.G., Verdicchio, M.: AI, Agency and Responsibility: The VW Fraud Case and Beyond, pp. 639–647. AI & Society (2019)
41. Santoro, M., Marino, D., Tamburrini, G.: Learning Robots interacting with humans: From epistemic risk to responsibility. AI Soc. **22**(3), 301–314 (2008)
42. van de Poel, I.: Embedding values in Artificial Intelligence (AI) systems. Mind. Mach. **30**(3), 385–409 (2020)
43. Alpaydin, E.: Machine Learning, Revised and Updated Edition. The MIT Press, Cambridge, Massachusetts (2021)
44. Johnson, D.G.: Technology with no human responsibility? J. Bus. Ethics. **127**(4), 707–715 (2015)
45. Razjigaev, A., Pandey, A.K., Howard, D., Roberts, J., Wo, L.: End-to-end design of Bespoke, Dexterous Snake-Like Surgical Robots: A Case Study with the RAVEN II. IEEE Trans. Robot. **38**(5), 2827–2840 (2022)
46. Powers, T.M., Ganascia, J.-G.: The Ethics of the Ethics of AI. in In: The Oxford Handbook of Ethics of AI, pp. 27–51. Oxford University Press, Oxford (2020)
47. Franssen, M., Kroes, P.: Sociotechnical Systems. In: Friis, J.K.B.O., Pedersen, S.A., Hendricks, V.F. (eds.) in A Companion to the Philosophy of Technology, pp. 223–226. Wiley-Blackwell, Malden, MA (2013)
48. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada, 2021. (2021)
49. Shneiderman, B., Rose, A.: Social Impact Statements: Engaging Public Participation in Information Technology Design, in *Proceedings of the Symposium on Computers and the Quality of Life*, (1996)
50. Friedman, B., Hendry, D.G.: Value Sensitive Design, Cambridge, Massaschusetts: The MIT Press, (2019)
51. Tubaro, P., Casilli, A.A., Coville, M.: The trainer, the Verifier, the imitator: Three ways in which human platform workers support Artificial Intelligence. Big Data Soc., **7**, 1, (2020)
52. Douglas, D.M., Lacey, J., Howard, D.: Ethical responsibility and computational design: Bespoke surgical tools as an instructive case study. Ethics Inf. Technol. **24**(1), 11 (2022)
53. Tomsett, R., Braines, D., Harborne, D., Preece, A., Chakraborty, S.: Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems, in *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, Stockholm, (2018)
54. McDermid, J.A., Jia, Y., Porter, Z., Habli, I.: Artificial Intelligence Explainability: The technical and ethical dimensions. Philosophical Trans. Royal Soc. A. **379**, 20200363 (2021)
55. Douglas, H.E.: Science, Policy, and the Value-Free Ideal. University of Pittsburgh, Pittsburgh, PA (2009)
56. Alexandra, A., Miller, S.: Ethics in Practice: Moral Theory and the Professions. University of New South Wales, Sydney (2009)
57. Boden, M.A.: GOFAI. In: Frankish, K., Ramsey, W.M. (eds.) in The Cambridge Handbook of Artificial Intelligence, pp. 89–107. Cambridge University Press, Cambridge (2014)
58. Domingos, P.: The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Penguin Books (2015)
59. Hermansson, H., Hansson, S.O.: A three-Party Model Tool for ethical risk analysis. Risk Manage. **9**, 129–144 (2007)
60. van de Poel, I., Nihlén, J., Fahlquist: Risk and responsibility. In: Roeser, S., Hillerbrand, R., Sandin, P., Peterson, M. (eds.) in Essentials of Risk Theory, pp. 107–143. Springer, Dordrecht (2013)
61. Klinke, A., Renn, O.: The coming of age of risk governance. Risk Anal. **41**(3), 544–557 (2019)
62. Malakar, Y., Lacey, J., Bertsch, P.M.: Towards responsible science and technology: How Nanotechnology Research and Development is shaping Risk Governance practices in Australia. Humanit. Social Sci. Commun. **9**, 17 (2022)
63. Douglas, D.M., Lacey, J., Howard, D.: Ethical risks of AI-Designed products: Bespoke Surgical Tools as a case study. AI Ethics. **3**(4), 1117–1133 (2023)
64. Wolff, J.: Five types of Risky Situation. Law Innov. Technol. **2**(2), 151–163 (2010)
65. Hawk, J., Microsoft Azure, A.I.: Data, and Application Innovations Help Your AI Ambitions Into Reality, 15 November 2023. [Online]. Available: https://azure.microsoft.com/en-us/blog/microsoft-azure-ai-data-and-application-innovations-help-turn-your-ai-ambitions-into-reality/. [Accessed 17 April 2024]
66. Reiss, J.: Public Goods, in *The Stanford Encyclopedia of Philosophy*, Fall E. N. Zalta, Ed., Metaphysics Research Lab, Stanford University, 2021. (2021) ed.
67. Lovett, F.: A General Theory of Domination and Justice. Oxford University Press, Oxford (2010)
68. Lovett, F.: Republicanism, (2022) https://plato.stanford.edu/archives/fall2022/entries/republicanism/
69. Maheshwari, K., Nyholm, S.: Dominating risk impositions. J. Ethics. **26**(4), 613–637 (2022)
70. Hansson, S.O.: The Ethics of Risk: Ethical Analysis in an Uncertain World. Palgrave Macmillan, Basingstoke, Hampshire (2013)
71. Brodkin, J.: After Robotaxi Dragged Pedestrian 20 Feet, Cruise Founder and CEO Resigns, 11 November 2023. [Online]. Available: https://arstechnica.com/tech-policy/2023/11/after-robotaxi-dragged-pedestrian-20-feet-cruise-founder-and-ceo-resigns/. [Accessed 17 April 2024]
72. Lewens, T. (ed.): Risk: Philosophical Perspectives, London and New York: Routledge, (2007)

73. Department of Industry, Science and Resources: Safe and Responsible AI in Australia: Discussion Paper, Australian Government, Canberra, (2023)
74. Guntzbunger, Y., Johnson, K.J., Martineau, J.T., Pauchant, T.C.: Professional ethnocentrism and ethical risk management efficacy: How engineer's emotional openness mediates this complex relationship. Saf. Sci. **109**, 27–35 (2018)