A Principle-based Ethical Assurance Argument for AI and Autonomous Systems

Zoë Porter, Ibrahim Habli, John McDermid

ABSTRACT

An assurance case presents a clear and defensible argument, supported by evidence, that a system will operate as intended in a particular context. Assurance cases often inform third party certification of a system. One emerging proposal within the ‘trustworthy AI’ and Autonomous Systems (AS) research community is to extend and apply the assurance case methodology to achieve justified confidence that a system will be ethically acceptable when used in a particular context. In this paper, we develop and further advance this proposal, in order to bring the idea of ethical assurance cases to life. First, we discuss the assurance case methodology and the Goal Structuring Notation (GSN), which is a graphical notation that is widely used to record and present assurance cases. Second, we describe four core ethical principles to guide the design and deployment of AI/AS: justice; beneficence; non-maleficence; and respect for personal autonomy. Third, we bring these two components together and structure an ethical assurance argument pattern - a reusable template for ethical assurance cases - on the basis of the four ethical principles. We call this a Principle-based Ethical Assurance Argument pattern. Throughout, we connect stages of the argument to examples of AI/AS applications and contexts. This helps to show the initial plausibility of the proposed methodology.

The aim of this paper is to start to shape the debate around ethical assurance cases. We welcome comments to the corresponding author.2

1. Introduction

Most interdisciplinary discussions about Artificial Intelligence (AI) start with a question about what the term ‘Artificial Intelligence’ actually means. What are we talking about when we speak of systems that embody AI? How does AI relate to Machine Learning (ML)? What exactly is an autonomous system?

Though many different interpretations may be given, we take the broadly functionalist perspective that AI is that which enables a machine to do what it takes intelligence for a human to do. And we work on the basis that this, in turn, encompasses the range of techniques listed in the definition given in the European Commission’s proposal for an

---

1 Assuring Autonomy International Programme, Department of Computer Science, University of York. Corresponding author: zoe.porter@york.ac.uk
2 We thank the following for their helpful comments on an earlier draft of this paper: Dr Christopher Burr; Professor Richard Clegg; Dr Kate Devitt; Dr Richard Hawkins; Dr Christopher Jay; Dr Tom Lawton; Ivan Kyambadde; Dr Colin Paterson.
Artificial Intelligence Act: machine learning (ML), including supervised, unsupervised, and reinforcement learning; logic and knowledge-based approaches, including knowledge representation and expert systems, inductive logic programming, inference engines, and symbolic reasoning; or statistical approaches, including Bayesian networks and search and optimisation techniques (European Commission 2021, Annexes). Although defining AI partly in terms of these techniques carries the risk of becoming outmoded, and also means that many systems that are now considered ‘traditional’ are included under the term ‘AI’, we adopt this definition because we want to take a broad view of AI, rather than identify it with any single technique, however dominant in current applications.

AI-enabled decision support systems (e.g. systems used for medical diagnostics), where it is the human-in-charge who actually implements the decision, we count as AI systems; they are therefore what we mean by the ‘AI’ bit of the ‘AI/AS’ descriptor. By ‘autonomous system’ (AS), we mean a software-intensive system that has a decision-making function which it executes independently of direct human control. Increasingly, autonomous systems embody AI, and specifically ML. At higher levels of autonomy, a whole complex task is delegated to an AS. At lower levels of autonomy, the system executes a part of the task independently of direct human control and then hands over to a human operator, or the human and AS share responsibility across a set of tasks being undertaken. AS may be physical platforms situated in a real-world environment (e.g. vehicles and submersibles), or they may be systems situated in purely digital environments (e.g. high-speed trading applications). In many cases, human performance of the tasks AS carry out would require not only expertise and situational awareness but also sensitivity to the ethical and legal dimensions of the decision and the environment (Burton et al. 2020).

Virtually every conceivable sector is affected by the development and application of AI/AS: agriculture; automotive; aviation; criminal justice; defence; education; energy; finance; healthcare; the humanitarian sector; insurance; manufacturing; maritime; nuclear; the police; retail; the sciences (physical, life, and earth); social care; space (United Nations 2021; Savage 2020; West and Allen 2018). The raft of consumer applications is also growing. The global annual value of private equity investments in AI start-ups grew from $3 billion in 2012 to nearly $75 billion in 2020, with companies developing driverless and mobility technologies attracting the most investment, followed by health care, drugs, and biotechnology (Tricot 2021). There has also been substantial public investment. Governments across the world – especially in the US, Canada, China, the EU, the UK, Japan, and Australia - are investing billions of dollars into AI/AS research and development (Savage 2020). The UK government, for example, has invested more than £2.3 billion into AI across a range of initiatives since 2014 (HM Government 2021).³

At the same time, the past five to ten years have witnessed intensive debates and discussions about the ethical and societal implications of AI and autonomous systems. Debates about these ethical implications led “seemingly every organisation with a connection

---

³ This figure excludes defence spend.
to technology policy … [to] author or endorse a set of ethical principles for AI/AS” (Fjeld et al. 2020: 4). In the past couple of years, however, the focus has been on bridging the gap between principles and practice. Some of the sets of principles are starting to translate into hard and soft regulation, such as the European Commission’s draft proposal for an Artificial Intelligence Act, published in April 2021, and the IEEE’s P7000 series of ‘ethical standards.’ And there is a wealth of activity around developing specific tools and processes to achieve ethical systems, from algorithmic auditing, to red-teaming, ethical black boxes, impact assessments, certification, and conformity assessment (CDEI 2021, Koshiyama et al. 2021).

This paper is situated within that shift towards the practical. But it takes the view that a core set of ethical principles should still be explicitly articulated within these practical efforts. We also propose that a specific kind of practical method – the assurance argument - can organise a complex web of inter-related ethical requirements very effectively, making it a valuable tool (or ‘tools of tools’) to achieve ethically acceptable AI/AS.

We define ‘ethically acceptable AI/AS’ as follows:

the design of the system, its outputs (i.e. its predictions, classifications, recommendations, manoeuvres), its use, and its outcomes (i.e. the consequences the outputs cause) align with, and remain aligned with, people’s reasonable expectations of fairness, and respect for their welfare as autonomous individuals, as well as for the welfare of society and the environment in which they live.4

This is a broad definition of ethical acceptability. It encompasses three levels of ethical concern (individual, societal, and environmental). It also requires that an ethical perspective is taken across the system's lifecycle. We also assume that, even if the users have good intentions, the use of a system cannot be satisfactorily ethical if the design of the system is unethical (while it is possible for a system to be designed ethically and misused). In the presentation of the argument later in this paper, therefore, we focus on use, motivated by the recognition that this is of the highest generality and perhaps the greatest value. To note, the goal is ethical acceptability not ethical perfection. Even if logically possible, ethical perfection is not a realistic goal for practitioners. But we suggest that the bar for ethical acceptability should be high.

The rest of the paper is structured as follows.

In section 2, we introduce the assurance case methodology, which is typically used to present safety cases for critical systems, and the Goal Structuring Notation (GSN), which is a graphical notation for the documentation and presentation of these cases.

---

4 The phrase ‘people’s expectations’ could be interpreted in either a descriptive sense (what people do expect) or a normative sense (what people should expect). We mean it in the normative sense.
In section 3, we present four core ethical principles, derived from the principles of medical ethics, that we believe apply to AI/AS. These principles are: justice; beneficence; non-maleficence; and respect for personal autonomy.

In section 4, we bring these two elements together and develop an assurance argument pattern, grounded in the four ethical principles and presented in GSN, that can be used to justify that a system will be acceptably ethical when deployed in its intended context. We call this Principle-based Ethical Assurance Argument pattern. The intention is that this argument pattern will provide a prototype for a reusable template for ethical assurance cases tailored to specific contexts, purposes, and applications of AI/AS.

In sections 5 and 6, we take a more detailed look at each component of the argument pattern. We take the justice argument (called the Justice Assurance Argument) to represent the ‘first principle amongst equals’ when determining the ethical acceptability of deployed systems. Feeding into the justice argument, and providing the claims and data to inform the evaluation of potential trade-offs within it and decisions about the equitable distribution of benefit and burdens, are the beneficence, non-maleficence, and respect for personal autonomy arguments (these are called the Beneficence Assurance Argument, the Non-Maleficence Assurance Argument, and the Personal Autonomy Assurance Argument, respectively).

In section 7, we conclude with a statement of the role this method could play given the current pressing question in deployment ethics: when would the decision to deploy or use a specific system be justified?

2. Assurance cases and GSN

‘Assurance’ refers to the general activity of providing justified confidence in a property of interest. Within engineering, that property of interest is, most commonly, safety. The assurance case methodology, specifically for safety, is typically used to “present a clear, comprehensive and defensible argument that a system is acceptably safe to operate within a particular context” (Kelly 1998: 3). It offers a proactive and evidence-based approach, in contrast to approaches that simply involve following a prescriptive process to justify confidence in the safety of a given system. Assurance cases are often required as part of the regulatory process. In highly-regulated sectors, such as defence, aviation and healthcare, they tend to inform pre-deployment certification of a system as safe to operate. Assurance cases are also required to be updated, given changes in the system and its environment, in order to maintain confidence in the actual behaviour and use of the system.

In addition to the fact that regulators often require an assurance case, they are widely used in safety-critical systems engineering for several reasons (Habli, Alexander and Hawkins 2021; The Health Foundation 2012). First, assurance cases are explicit. They make it clear why it would be reasonable to believe that the claim regarding the property of interest is true, without declaring that deductive proof of the truth of the claim has been given. This provides
a good basis for critical scrutiny and discussion about the claims made by those who put a system forward for deployment. Explicitness also provides a practical basis for assurance cases to be revised iteratively to reflect new evidence and learning. Second, by breaking down claims in the way that they do, assurance cases promote structured and methodical reasoning amongst stakeholders, including decision-makers. Third, they function as a ‘tool of tools,’ able to incorporate several other assurance mechanisms (e.g. impact assessments and audits) within their overall structure. The justificatory and consolidatory role of assurance cases has made them a primary requirement for demonstrating compliance with national and international standards (e.g. ISO 2011). Fourth, assurance cases provide a framework that enables the integration of evidence from multiple sources. Fifth, because of their explicitness, their clarity, and their dialogical form, assurance cases can foster interdisciplinary and multidisciplinary collaboration.5

Assurance cases can be presented using different notations. One such notation is the Goal-Structuring Notation (GSN), which was developed at the University of York in the 1990s (Kelly 1998; McDermid 1994). Based on a model of argumentation promoted by Stephen Toulmin (1958), it places an emphasis on informal, yet well-structured, justification. Assurance cases structured in accordance with GSN are hierarchically decomposing, arguing back from a top goal - the claim that the argument supports - via an argument strategy which elucidates the inference that exists between a goal and the sub-goals that support it. These sub-goals in turn are supported by evidence and data. The evidence and the argument strategy together provide abductive support for the claim expressed in the top-level goal. As Goodenough and colleagues put it, “an assurance case provides defeasible reasons for believing that a claim is true” (Goodenough et al. 2012: 27). The main symbols and elements of a GSN assurance argument are presented in Figure 1 below6.

5 To clarify, we understand ‘interdisciplinary’ to mean the integration and synthesis of different disciplinary perspectives in the final research or policy output, and ‘multidisciplinary’ to mean the drawing upon different disciplines which nonetheless stay within their boundaries (Choi and Pak 2006). As an example, this paper puts forward an interdisciplinary proposal.

6 The reader is advised to consult the Goal Structuring Notation Community Standard for a detailed explanation of GSN and its associated methodology (Assurance Case Working Group 2021).
Figure 1 Symbols and elements of a GSN argument, extracted and adapted from Assurance Case Working Group (2021)

3. Ethical principles

The last few years of the 2010s witnessed a proliferation of ethics declarations for the development and use of AI and autonomous systems, with well over 80 major sets of ethical principles published by intergovernmental and national government agencies, public bodies, NGOs, corporations, professional institutes and universities worldwide (Fjeld et al. 2020; Jobin, Lenca and Vayena 2019). From meta-reviews of these ethical principles, several scholars have noticed a striking overlap between the core themes that have emerged for the ethics of AI/AS and the principles of medical ethics (Morley et al. 2020; Floridi and Cowls 2019; Floridi et al. 2018). This finding has been endorsed by the OECD, amongst others (Mittlestadt 2019).

The four classical principles of medical ethics are:

- **The principle of respect for autonomy**, which requires that people are free from controlling interference and can live according to their own reasons and motives
- **The principle of non-maleficence**, which imposes an obligation not to inflict (unjustified) harm
- **The principle of beneficence**, which requires positively providing benefits to others and contributing to their welfare
- **The principle of justice**, which requires that benefits and burdens are equitably distributed

These four ethical principles were first articulated and presented by Beauchamp and Childress in 1979 (Beauchamp and Childress 1979). They quickly became dominant in the field, transforming the way medical ethics was understood and practised. They are taught in almost every medical school in the world. The approach to medical ethics that is grounded in this core set of principles has come to be known as ‘principlism’ or the ‘four-principles approach’. Ethical principles are ethical values expressed in normative form. They can play an action-guiding role: they provide reasons that favour some actions; and they can be used to justify those actions to other people. The four-principles approach was motivated by the need to provide an ethical framework for guiding action in cases where different values are at stake in medical dilemmas, without either over-simplifying the complexity of those dilemmas or getting immersed in intractable disagreements at the level of ethical theory (Holland 2015).

We seek to advance the position that these four principles of medical ethics have a clear relevance to the ethics of AI/AS - so long as they are suitably adapted and adjusted to this

---

7 The principle of beneficence is not generally understood unrestrictedly. Typically, it applies to those parties with some special responsibility for others, for example because it falls within their duty of care to advance the welfare of others. We thank Dr Chris Jay for highlighting this, and we discuss it in section 5.2.
new ethical domain. Moreover, and to clarify, we seek to advance the position that this relevance is not limited to systems used in healthcare domains.

There are fundamental similarities between medicine and the engineering and deployment of AI/AS that support this position. Both medicine and engineering are applied sciences. They both involve decision-making that affects peoples’ lives and wellbeing in complex real-world environments. In both cases, risks and benefits need to be carefully weighed. Those often incommensurable risks and benefits also have a wider scope than the immediate risk-bearer or beneficiary. Trade-offs are an important part of medical ethics (e.g. when balancing the benefits of a high-risk medical intervention for one patient against the resources it would take from elsewhere and from other patients). Trade-offs will also inevitably arise when making decisions about the development or deployment of an AI/AS. One clear advantage of principlism is that principles are defeasible - in specific cases, and in the light of specific considerations, new evidence or disclosures, the requirement of beneficence, for example, may be revised or rejected because the obligation not to impose harm is overriding strongly; each principle may be defeated by the stronger reasons provided by another principle in the specific context. This makes a principle-based approach appropriate for a framework in which trade-offs will almost certainly need to be made. Another deep alignment between medical ethics and the ethics of AI/AS concerns justice: just as health inequalities across demographic groups are a pervasive global concern, so too is the risk of entrenched inequalities from the deployment of AI and autonomous systems. And in both cases, the autonomy of the immediately affected individual is at risk of being constrained. These reasons ground our proposal that the four widely accepted principles of medical ethics are appropriate here and can provide coverage of the central ethical concerns that the use of AI/AS raise.

The adaptation of the principles to the ethics of AI/AS is critical, however, because despite fundamental similarities, there are also differences between medical and engineering ethics. Medical practice is about health and healing. Engineering and technology is about innovation and problem-solving, and is more closely tied to commercial imperatives. The relation between clinicians and their patients is more personal than the relation between engineers (and other AI/AS actors) and users. Clinicians have duties of care towards patients that do not map easily onto the professional responsibilities of engineers (Mittlestadt 2019). In medical practice, the authoritative decision-makers (historically) have been, exclusively, human beings; in the ethics of AI/AS, a new kind of ‘decision-maker’, the system itself, is brought into the causal chain, with a qualitatively different kind of ‘autonomy’. To be clear, the AI/AS does embody the values of the human decision-makers who contributed to its development, and so its autonomy cannot be detached from human values. But a difference remains: the system is a proxy for human beings in a way in which individual clinicians, as moral agents with minds of their own, are not simply proxies for everyone by whom they have been taught or influenced. In section 5 of this paper, we discuss in more detail how each ethical principle could be adapted and adjusted to the AI/AS ethics domain.

---

8 We thank Dr Kate Devitt for raising this point.
A group of academics affiliated with the Oxford Internet Institute have also advocated basing practical ethical methods for AI/AS around these four ethical principles - and, in this sense, the proposal we advance is not new (Morley et al. 2020; Floridi and Cowls 2021; Floridi et al. 2018). But, aside from the specific practical assurance method that we advocate, our account diverges from these authors in two respects. First, they recommend the addition of a fifth principle, ‘explicability’ - a combination of explainability and accountability (Floridi and Cowls 2021: 5-6). Our approach is to present ‘transparency’ (both as a property of the system and as a property of human decision-making about and around the system) as orthogonal to the original four principles; and present it as a sub-argument which, when the transparency is of appropriate kind and degree, provides an enabling condition for the enactment of the other four principles within the overall assurance argument (Turilli and Floridi 2009). This is discussed in section 6 of this paper. Second, we privilege the ethical principle of justice as the ‘first principle amongst equals’ in the context of the ethics of AI/AS. The principle of justice takes methodological priority because the Justice Assurance Argument is the site of trade-offs between the values advanced by the Beneficence Assurance Argument, the Non-maleficence Assurance Argument, and the Personal Autonomy Assurance Argument. The principal requirement for ethically acceptable uses of AI/AS, we suggest, is that benefits, risks of harm, and impacts on personal autonomy are evaluated and distributed equitably. This is discussed in section 5.1 of this paper.

The four principles plausibly offer a unifying normative basis upon which to proceed when assuring systems as ethically acceptable for use in their intended contexts. This is largely because they represent a distillation of the core themes that have emerged from major sets of ethical principles, which in turn have informed regulatory developments. Most of the regulations, guidelines, and standards that already exist can be subsumed under the aegis of the four principles. Floridi and Cowls offer several examples (Floridi and Cowls 2021). Here, we offer two to illustrate our point. One is the European Commission’s proposal for the world’s first legal framework for the governance of AI (European Commission 2021). This proposal emphasises safety and the fundamental rights and freedoms of citizens. These requirements can be understood as falling under the principles of non-maleficence (safety), justice and non-maleficence (rights), and respect for autonomy (freedoms), respectively.

Another example is the report of the UK’s House of Lords Select Committee on Artificial Intelligence, which recommended a cross-sectoral ‘AI code’ that could provide the basis for statutory regulation, “if and when this is determined to be necessary” (House of Lords 2018: 125). The Committee proposed five overarching principles around which such a code could be built:

(1) Artificial intelligence should be developed for the common good and benefit of humanity.
(2) Artificial intelligence should operate on principles of intelligibility and fairness.
(3) Artificial intelligence should not be used to diminish the data rights or privacy of individuals, families or communities.
(4) All citizens have the right to be educated to enable them to flourish mentally, emotionally and economically alongside artificial intelligence.
(5) The autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence.

These can be understood as falling under the principles of beneficence (1, 4), non-maleficence (3, 5), respect for autonomy (3, 5), and justice (2, 3), respectively.

There are further pragmatic reasons to adopt the four-principles approach. Their usability and memorability has been tried and tested in the medical field. They have the potential to be an accessible heuristic for a wide range of non-philosophical professionals, including designers, engineers, safety teams, manufacturers, operators, and users - their function as such has been shown in the field of medical ethics. Principlism has several critics amongst academic bioethicists but in practice, clinicians find them helpful to reason about medical dilemmas (Gillon 1994) - something our clinical colleagues support through anecdotal evidence. For all of the reasons given in this section, our position is that the four ethical principles provide a plausible basis to achieve ethically acceptable uses of AI and autonomous systems.

4. An ethical assurance argument based on the four ethical principles

The idea of developing an argument-based ethical assurance case has received its most sustained support in a recent paper by Burr and Leslie (2021), (see also: Hauer, Adler and Zweig 2021; ICO/Turing 2020, Annexe 5; Menon and Alexander 2020). Described by Burr and Leslie as the “process of using structured argumentation to provide reviewable assurance that a particular set of normative claims about the corresponding ethical properties of a system are warranted given the available evidence,” argument-based ethical assurance involves taking the methodology historically used to achieve and communicate confidence in the safety of systems and adapting it to a more broadly ethical goal (Burr and Leslie 2021: 17).

The aforementioned advantages of assurance cases are particularly salient in the context of ethical assurance. The focal role of argument in the assurance case methodology is appropriate for ethical dilemmas and decision-making, where there may be several different plausible answers to the same question. More precisely, the method is based on informal logic and abductive reasoning: a type of reasoning that builds confidence in the truth of a claim by way of inference from the best justification. This, too, is appropriate for ethical decision-making, where answers are not certain. Moreover, explicit reasoning that can be effectively and rigorously scrutinised is critical if the use of the systems is going to be ethically acceptable over the long term. Scrutiny aids discussion, open debate, and continued improvement. That the method is a ‘tool of tools’ is a particular strength given the

---

9 This is, therefore, ‘ethical assurance’ narrowly construed to refer to ethical assurance arguments and cases, rather than broadly construed to refer to the whole ecosystem of tools and services that can be used to achieve ethical AI/AS (e.g. CDEI 2021).
vast number of tools or mechanisms that are being proposed to deal with the ethical implications of AI/AS and the increasing need to organise all of the material in practicable ways (CDEI 2021). Furthermore, interdisciplinarity and multidisciplinarity is key to effective solutions in the complex arena of ethical AI/AS, which involves not just design and engineering practice, but the involvement of human factors experts, as well as the perspectives of ethicists, regulators, social scientists, and lawyers.

In this paper, we start substantially to develop the idea of adapting the assurance case methodology beyond safety. More specifically, we present a Principle-based Ethical Assurance Argument pattern, which we present in GSN. Argument patterns are “reusable templates” for assurance cases (Kelly and McDermid 1997). We call it a ‘Principle-based Ethical Assurance Argument pattern’ because the argument strategy is to support the claim of ethical acceptability by demonstrating enactment of the four ethical principles discussed in section 3. To orient the reader, Figure 2 below shows the overall structure of our argument pattern. This gives its highest-level presentation. It presents the argument as a composition of modules, which encapsulate the separate but interdependent argument structures that comprise the overall argument. The rationale for introducing the argument by way of this modular presentation is to give the reader an ‘at-a-glance-view’ of how the overall argument fits together, before describing its individual parts in more detail.

Figure 2 Modular structure of the argument
This argument pattern can be described here in a little more detail. The Principle-based Ethical Assurance Argument module at the top is where the overall argument’s highest-level goal is contained. More precisely, as we describe below in Figure 4, this goal is the well-justified claim that, for the intended purpose, the use of the system will be ethically acceptable in the intended context. The Principle-based Ethical Assurance Argument module is immediately supported by the Justice Assurance Argument module, which contains its own top-level goal (see Section 5.1). The Justice Assurance Argument module is, in turn, directly and conjointly supported by the Beneficence Assurance Argument module, the Non-maleficence Assurance Argument module, and the Personal Autonomy Assurance Argument module - which each contain their own top-level goals (see Sections 5.2 - 5.4). The purpose of the Transparency Assurance Argument module is to show that the relevant transparency requirements for each of the four principles have been met (see Section 6 for discussion).

Adjacent to the Principle-based Ethical Assurance argument module is the Principlism Confidence Argument module. This module is inspired by the idea, introduced by Hawkins and colleagues, of separating out the main safety argument and the confidence argument (Hawkins et al. 2011). The confidence argument offers a more expressive and more sustained justification for a particular part of the argument than the use of the justification symbol and element in GSN (see Figure 1). We employ this approach to justify the principle-based approach that grounds the overall argument, and specifically to justify the approach of using the four principles adapted from medical ethics: justice; beneficence; non-maleficence; and respect for personal autonomy. The Principlism Confidence Argument is presented in Figure 3. It organises and summarises much of the discussion in section 2 of this paper.

The top-level goal contained within the Principlism Confidence Argument module is the well-justified claim that the four ethical principles provide a plausible normative basis for achieving the ethically acceptable use of an AI/AS (PG1). The argument strategy - which describes the inference that exists between this goal and the sub-goals - is that confidence will be justified if the four-principle approach is shown to be appropriate, adequate, and pragmatically advantageous (PA1). This, in essence, is what was discussed in section 2. The appropriateness (PG2) is justified by the claims that: a framework based on principles allows for trade-offs (PG5); there are sufficient relevant similarities between the original context in which these four principles were proposed and the context of the ethics of AI/AS (PG6); and the four principles distil the main themes that have emerged from the meta-reviews of published sets of ethical principles (PG7). The adequacy (PG3) is justified by the claims that: many other proposed ethical principles can be shown to be derivative upon the core four (PG8); and it is not clear that important ethical values would be lost by restricting the grounding of this argument to the four principles (PG9). Finally, the pragmatic advantage (PG4) is justified by the claims that: the principles are an accessible heuristic for a wide range of practitioners (PG10); and they look initially promising for meeting emerging regulatory requirements across a diverse international regulation and governance landscape (PG11).
Figure 3 Principlism Confidence Argument
Table 1 (PS2) provides the evidence to support the claim at (PG8). This table shows that most of the other proposed principles such as sustainability, safety, and fairness, as highlighted in meta-reviews of the principles (Fjeld et al. 2020; Jobin, Ienca and Vayena 2019) are actually derivative upon the four core principles of justice, beneficence, non-maleficence, and personal autonomy. They can be covered by the core set of principles, because they express values that fall within the values expressed by those principles. It should be recognised that not all of the values expressed by the principles fit neatly under one principle alone. Invasions to privacy, for example, are harms, threats to personal autonomy, and injustices (and against some laws, e.g. General Data Protection Regulation (GDPR) and UK Data Protection Act 2018) - a principle of privacy, therefore, sits under non-maleficence, personal autonomy, and justice.

It seems to us that ethical principles which have been highlighted in the meta-reviews and which are not clearly derivative upon the four principles (i.e. transparency, trustworthiness, solidarity, and respect for human values) are connected to those principles in ways that do not undermine the claim in (PG8). This relation is also captured in Table 1.

| Principle | Beneficence | Non-maleficence | Autonomy | Justice |
|-----------|-------------|-----------------|----------|---------|
| Covers    | Flourishing | Safety          | Human control | Fairness |
|           | Sustainability | Security       | Dignity    | Accountability |
|           |              | Privacy         | Privacy    | Responsibility |
|           |              | Non-discrimination |              | Reciprocity |
|           |              |                 | Privacy    | Privacy |
|           |              |                 | Non-discrimination | |
| Successful | Transparency | Transparency | Transparency | Transparency |
| enactment of |             |                |            |         |
| principle is enabled by |            |                |            |         |
| Consequence | Trustworthiness | Respect for Human Values | Solidarity |         |
| of successful | (in particular, as a consequence of justice) |            |            |         |
| enactment of |              |                |            |         |
| the principles |              |                |            |         |

Table 1 Coverage of the ethical principles proposed for AI/AS

The proposed principle of transparency, we maintain, is a ‘pro-ethical’ condition or value that, when appropriately expressed in action, in fact enables the successful enactment of the four core principles.\textsuperscript{10} It is important to emphasise the need for caution here, hence the stress on the word ‘appropriately’: some forms and approaches to transparency can be

\textsuperscript{10} To reiterate, we mean transparency both in the sense of ‘technical transparency’ (or ‘explainable AI’) and ‘governance transparency’ (i.e. visibility of human decision-making about and around the system over the lifecycle).
counter-productive to ethical acceptability. For example, excess transparency may in fact occlude or distract attention away from a developer’s intent or decisions of significance, and it may also be used to undermine a user’s personal autonomy (Ananny and Crawford 2018).\(^{11}\) Moreover, transparency is inimical to some security concerns. The equivocal role that transparency can play provides reason not to include it as a separate ethical principle within the Principle-based Ethical Assurance Argument pattern. Transparency is not a desideratum as an end in itself; within the context of this argument, it is a desideratum only insofar as it enables the use of AI/AS to be beneficial, non-harmful, respectful of personal autonomy, and in accordance with the demands of justice.

We further suggest that other proposed principles that are not clearly derivative upon the core four - trustworthiness and ‘respect for human values’ (Fjeld et al. 2020) - would in fact be better understood as ethically valuable states of affairs that are either yielded (in the case of trustworthiness) or demonstrated (in the case of respect for human values) by the successful enactment of the four core principles. Another principle that has been identified in the meta-reviews is the principle of solidarity (Jobin et al. 2019). It seems to us that solidarity - a sense of unity, mutual attachment, and commitment to the common good - can also be understood as an ethically valuable state of affairs which would be consequential upon the successful enactment of the principle of justice, in particular.\(^{12}\) Finally, while we believe that Table 1 plausibly supports (PG8), we do also recognise that future ethical principles that might be proposed for AI/AS, or ones we have not considered, might not map onto this framework.

Having set out the modular structure of the overall argument pattern, and having given the Principlism Confidence Argument, we can now look - in Figure 4 - more closely at the highest-level claim of the overall Principle-based Ethical Assurance Argument pattern.

The highest-level goal of the overall Principle-based Ethical Assurance Argument pattern is to achieve justified confidence that in the truth of the claim that, for the intended purpose, the use of the system will be ethically acceptable within the intended context (TG1). This goal is situated within a series of explicit contextual definitions or statements, called ‘contextual artefacts’ within GSN (TC1-TC5).\(^{13}\)

\(^{11}\) With thanks to Dr Chris Burr for bringing this point - and this paper - to our attention.

\(^{12}\) To note, Rawls, by contrast, claims that solidarity plays a role in underwriting principles of justice (1971: 90-91).

\(^{13}\) To note, we have not included a list of stakeholders here as a contextual artefact - three groups of stakeholder are instead described in the sub-arguments below (in section 5.2, 5.3 and 5.4).
Let us now consider how those contextual artefacts would be instantiated. The definition of ‘ethically acceptable’ (TC1) has been given in the Introduction of this paper. The others will be described through two examples, which will be a running theme throughout the rest of the paper - and which, we hope, will start to bring this argument to life. The first example is the hypothetical example of an autonomous ‘robo-taxi.’ The definition of the intended purpose might be that the robo-taxi should transport passengers without a human taxi driver in the vehicle within a specified operating environment or operational design domain (ODD) (TC2). The definition of the usage might be that the vehicles are used to take passengers from a city’s major railway station to various locations within that ODD, which covers a 5-mile radius in the city (TC3). The definition of the system would be its description, as a physically embodied system of systems, with a sense-understand-decide-act (or ‘SUDA’ loop) design pattern, and a description of the computational techniques used to build it, and so on (TC4). And the intended context would be the environment in which it is used, including details such as: the position and relevant features of the particular railway station; the particular city; its weather conditions; relevant features of the population; typical traffic and pedestrian flow; and the precise limits of the ODD (TC5).
The second example we give is of a clinical AI system. This example is not hypothetical, but an existing system called the ‘AI Clinician’ (Komorowski et al. 2018). The system’s intended purpose is to make recommendations on one of two of the many possible treatments of sepsis (vasopressors or intravenous fluids) for adult patients with a diagnosis of sepsis on an Intensive Care Unit (ICU) ward (TC2). The definition of the usage is that it is used to recommend one of these two treatments for the patient over the course of a four-hour window (TC3). The definition of the system would be its description, as a digital system built using reinforcement learning techniques, which analyses 48 features from an electronic patient record and uses a Markov decision process to simulate clinical decision-making (TC4). And the intended context is a busy ICU ward in an NHS teaching hospital (TC5).

The argument strategy (TA1) is that, by demonstrating that each of the four ethical principles have been enacted, developers or operators putting such systems forward will be able to justify confidence in the argument’s highest-level claim that, for the intended purpose, the use of the systems will be ethically acceptable in the intended context (TG1). This strategy requires a further contextual artefact, namely a description of each of the four ethical principles of justice, beneficence, non-maleficence, and respect for personal autonomy (TC6). This description was given in section 3 of this paper. The argument strategy is supported by the Principlism Confidence Argument (see Figure 3), presented here as an ‘away goal’ to denote the fact that it repeats a claim earlier represented in an argument module.

5. Goal decomposition and the four principle-based assurance arguments

Let us now consider each of the four principle-based assurance arguments that conjointly justify the overall Principle-based Ethical Assurance argument’s highest-level claim of the ethical acceptability of the use of the system.

5.1 The Justice Assurance Argument

The breadth and depth of ethical concerns about the use of AI/AS reveals that there is a need for an enlarged ‘trade space’ (the range or scope of anticipated impacts from the use of a system) within which to reason about the acceptability of AI/AS, beyond a narrow focus on reducing risk to a level that is ‘as low as reasonably practicable’ (McDermid, Porter and Yia 2021). Our hope is that the Justice Assurance Argument, as supported by the other three principle-based sub-arguments, fulfils that enlarged ‘trade space’ function.

We have said that the Justice Assurance Argument takes methodological priority over the other three. The reason for this is that there will very likely be a range of comparisons and trade-offs to be made between the benefits, residual risks of harm, and impacts on personal

---

14 ‘As low as reasonably practicable’, or ALARP, is a principle in the regulation of safety-critical systems that derives from the Health and Safety at Work Act 1974. Determining that risks have been reduced ALARP generally involves an assessment of the risk in question, the money, time and effort that would be involved in taking measures to avoid it, and a comparison of the two.

15 From this point in the paper, though we are still talking about argument modules, we drop the word ‘module’ for ease of readability.
autonomy that use of the system will give rise to - and the equity or fairness of the distribution of these, all things considered, is what will render the decision to deploy or to use the system in context defensible. Ultimately, it is the justification of this distribution that will support the highest-level claim of ethical acceptability. Questions about equitable distributions of benefit and burden fall under the aegis of the principle of justice.

The central idea here is that each of the other three principle-based assurance arguments function as precursors to the Justice Assurance Argument, by providing the information and partial justifications necessary to start to compare different distribution options and decide upon that which is most ethically defensible (see Figure 5). The diamond symbol beneath the justice claim ($JG1$) indicates that the argument supporting this claim is partially developed. That is, for a complete argument, the justice claim ($JG1$) will be supported by the beneficence ($BG1$), non-maleficence ($NG1$) and personal autonomy ($AG1$) claims in addition to claims that are specific to the justice principle (e.g. addressing the risk-benefit-autonomy trade-offs and determining their equitable distribution).

![Figure 5 Justice Assurance Argument](image)

More precisely, the Beneficence Assurance Argument, shown in Figure 6, provides information about, and partial justification of, what the realised or realisable benefits of the system are, and for which beneficiaries. The intent is that this will yield a benefit matrix which feeds into the Justice Assurance Argument. The Non-maleficence Assurance Argument, shown in Figure 7, provides information about, and partial justification of what the system’s managed and residual risks are, and for which risk-bearers. This, in turn, will yield a residual risk matrix which feeds into the Justice Assurance Argument. The Personal Autonomy Assurance Argument, shown in Figure 8, provides information about, and justification of, how the system may constrain the personal autonomy of individuals in context, and for which bearers of this particular kind of autonomy-undermining risk. This will yield a personal autonomy matrix which feeds into the Justice Assurance Argument.
These three informational elements - these three matrices - should then make clear what are
the different benefit-risk-autonomy distribution options, the interdependencies between them,
and the trade-offs that they incur. For example - and this is very much a speculative example
to illustrate the idea - in the hypothetical robo-taxi case, convenience to and novelty for end
users, as well as economic benefits to the local council and to the system’s developers,
might need to be compared against the risk of loss of earnings for local taxi drivers, which
might also need to be compared against the personal choice of end users, and any
constraints on the personal autonomy of residents in the ODD. Justification for the
highest-level goal of the overall argument (JG1) becomes complete once the reasoning
within the Justice Assurance Argument has been completed.

The demands of ethical acceptability mean that it is not enough simply to weigh risk-severity
and risk-probability when deciding whether it would be fair and defensible to use an AI/AS.
This is true even when the concept of risk is enlarged to include non-physical harms (as is
described in section 5.3 below). The personal autonomy and agency (under which category
we include personal choice, amongst other things) of the risk-bearers need to be included
too (Hansson 2003, 2018). It is for this reason that we include personal autonomy as an
element to be incorporated and evaluated in the Justice Assurance Argument. Moreover,
building on the work of Hansson (2018), we think there is merit to considering the
combinations of roles when determining the distribution of benefits and burdens. As a
general rule, it would be inequitable to use a system in which the risk-bearers are not also
amongst the beneficiaries of its use, for example.

Of all of the arguments, the Justice Assurance Argument is the one where we, the authors,
face the greatest number of as yet unresolved research questions. For this reason, we do not
give an initial decomposition of this argument in GSN, represented as an argument module,
as we do with the other three principle-based supporting arguments (in sections 5.2, 5.3, and
5.4 below). However, though this part of the overall argument is still undeveloped, we expect
that the top goal of the Justice Assurance Argument (‘the distribution of the benefit and risk
(of harm) is equitable, given the principle of autonomy’) will decompose into the immediate
sub-goal that ‘the use of the {system} represents a fair or equitable trade-off for all affected
stakeholders’.

We also propose that a decision procedure or method known as ‘reflective equilibrium’
should be employed to reason about the trade-offs. This method is most closely associated
with the work of the political philosopher John Rawls, who coined the term (Rawls 1951;
Rawls 1971). On our understanding of it as an actionable decision procedure or method that
will be employed in the Justice Assurance Argument, it will involve working back and forth
between: a) ethical judgements or intuitions about the specific AI/AS in particular intended
contexts; b) the four ethical principles that, we argue, should govern development and
deployment decisions more generally; and c) relevant non-ethical beliefs and judgements
(e.g. technical, legal, pragmatic) that may need to be factored into the decision. The inclusion
of non-ethical judgements within the decision procedure renders it a Wide Reflective
Equilibrium (Daniels 1979). The aim is to reach a decision that represents an acceptable
coherence among these judgements, principles, and beliefs. Reflective equilibrium is reached when we are no longer inclined to revise any of them, or their relative strength or weight, because together they have the highest degree of acceptability. Our working position is that the benefit-risk-autonomy distribution that is agreed upon by this procedure would be the most defensible basis upon which to decide to deploy a given system. Naturally, the procedure might also yield the conclusion that a particular AI/AS should not be used in the intended context at all.

There are several reasons that favour the Wide Reflective Equilibrium approach. First, it enables different kinds of affected stakeholder to be included in the decision procedure, such as engineers, management, service providers, ethicists, lawyers, social scientists, regulatory advisers, as well as - crucially - the people who would be directly affected by use of the system (e.g. expert users, current practitioners, local residents, and patient representatives). Of course, sometimes representatives or proxies for members of the public, such as regulators, will instead be involved. This participatory decision-making upholds the central ethos of Responsible Research and Innovation (Owen and Stilgoe 2013) and provides a promising basis for justifying the decision to deploy to the wider public. Second, because of the decision procedure’s emphasis on ethical and non-ethical judgements and beliefs about particular cases in particular contexts, and not just a focus on overarching ethical principles, the decision reached by the method should be meaningfully context-sensitive. Third, it allows for a consideration of options for action that involve incommensurable values - values that cannot be reduced to a common measure or to straightforward cardinal comparison, such as comparisons or trade-offs between system safety and user privacy, or between system performance (and hence, for example, improved clinical outcomes) and the meaningful control of the expert user (such as the clinician). Finally, reflective equilibrium is a valuable method for making decisions where there is uncertainty: this uncertainty can be included as a feature to be considered when making ethical judgements about particular cases, and where the risk matrices cannot be simply understood as denoting probability and severity of risk, since there is insufficient confidence in the probability of the risk.16

A question that, we think, demands open debate is that of the scope of the enlarged ‘trade space’ of the Justice Assurance Argument. When considering the distribution of benefits, risks, and constraints on personal autonomy, how far upstream should we go? Should we include all possible risk-bearers, for example, all the way to the very earliest stages of the supply chain, back to the working conditions of those mining the minerals from the ground, or to the factories in which the semiconductors are produced, or where the pixel-precise labelling of images for ML training datasets takes place? These are not trivial questions but questions with far-reaching geo-political implications. By the same token, how far downstream should we go? Clearly, the scope needs to extend to users and those in the immediate operating environment, but what about individuals in ten years' time, or future

16 As we further develop this work, we will consider also how to frame the equilibrium so that it holds across a period of time, where there may be several deployment or system updates, such that a whole new process of Wide Reflective Equilibrium need not take place, so long as the capabilities of the proposed system releases were encompassed within the exercise of the original decision procedure.
generations who may be positively or negatively affected by the use of the system? For example, widespread deployment of even beneficial AI/AS may have negative consequences on the environment, and impose increased risk on future generations (McDermid, Porter and Yia 2022). The temptation is to keep the model manageable by restricting scope, but the danger is that this becomes a form of willful myopia and systems that assimilate or threaten serious harm to people still pass the official threshold of ethical acceptability. If the ultimate motivation for the development and deployment of these technologies is to benefit humanity, the scope of the Justice Assurance Argument matters radically.

5.2. The Beneficence Assurance Argument

The promise that AI/AS can bring benefits to humanity is often the very reason given for their development and deployment (HM Government 2021). Within the ethically-aligned engineering community, prioritising human well-being is the principal desideratum of system design (IEEE 2018). The principle of beneficence expresses the central requirement that “there should be a (sought-after) benefit of having the system in the first place” (McDermid, Porter and Jia 2021: 4). The ethical principle of beneficence requires positively benefitting people, and specifically to contribute to their welfare or ‘well-being’, by which we mean the degree to which they have the essential goods or capabilities that make life go well for them (Crisp 2021). Considered in the long-term, it is not just the welfare of individuals but also of society as a whole and of the environment that matter.

In the context of medical ethics, Beauchamp and Childress claim that promoting the welfare of patients - and not merely avoiding harm - embodies the goal, rationale, and justification of medicine itself (Beauchamp and Childress 2009: 205). In one respect, this is true also of engineering, insofar as technology is often built to advance human capacities. But a tension inevitably arises that systems are often developed and manufactured by for-profit corporations, and it is far from established the degree to which an obligation of beneficence applies here (Beauchamp 2019). Indeed, this may explain why, within the sets of ethical principles for AI/AS, “references to non-maleficence outweigh those to beneficence by a factor of 1.5” (Jobin, Lenca and Vayena 2019). As mentioned in footnote 6 of this paper, the principle of beneficence generally applies to those who have a special responsibility towards others. An important and under-discussed question is whether, or to what degree, engineers, manufacturers, and operators have such a responsibility positively to advance the welfare of those people who are directly affected by AI/AS (as well as the welfare of society and the environment), and not just avoid imposing a risk of harm.17 It would be morally praiseworthy if they considered positive contributions to welfare in this way (and it would also be consistent with many of the narratives and motives such actors tend to articulate when communicating the value of having the system in the first place).

17 The Engineering Council and Royal Academy of Engineering’s joint statement of ethical principles offers an interesting view into this question; though principally focused on non-maleficence and professional integrity, it does also hold that engineering professionals have a duty to maximise the public good (Engineering Council and Royal Academy of Engineering 2017: 2)
More importantly, we take it that regulators and public bodies - who should ultimately decide whether the use of an AI/AS will be sufficiently ethically acceptable to warrant approval for deployment - do have such a special responsibility towards those affected by the system. If an ethical assurance case becomes a pre-deployment requirement for some AI/AS, it is wholly appropriate to the function and aims of such public institutions that benefits to welfare are documented and realised, and that the beneficiaries are clearly identified. This is what the Beneficence Assurance Argument, presented in Figure 6, provides.

The top-level claim of the Beneficence Assurance Argument (BG1) is that the use of the system provides benefits to, or is beneficial to, identified groups of beneficiaries. To illustrate how the Beneficence Assurance Argument could work in practice, we can return to the two examples. From the use of the hypothetical robo-taxi, the groups of possible beneficiaries, indicated at (BC2), include, at the individual level: i) the end users of the service; ii) other individuals in the ODD, such as pedestrians and local residents; and at the group or societal level: iii) the service-provider or operator of the robo-taxis; iv) the municipal or city council; and v) the developer or manufacturer of the system. The kinds of benefit use of the system

18 To recall the discussion in the final paragraph of section 5.1, it is still an important open question, however, quite how far this special responsibility extends.
might realise, indicated at (BC3), will inevitably be different for each of these beneficiaries. Plausibly, the benefit to end users is convenience and efficiency, and possibly also safety or mobility benefits, and the benefit to other individuals in the ODD may be increased safety and environmental benefits from quiet and energy-efficient systems. The benefit to the service-provider, as well as to the developer or manufacturer, will most likely be financial (and this may indirectly also be a wider societal or economic benefit). The benefit to the local municipal council might stem from improved efficiency and traffic flow, or from greater safety, or from reduced costs.

From the use of the AI Clinician, the groups of possible beneficiaries, which again would furnish the content of (BC2), include, at the individual level: i) the patient; ii) the individual clinician in-charge; iii) other clinicians and patients in the operating environment; and at the group or societal level: iv) the hospital or NHS trust; v) the developer or manufacturer of the system. With respect to the kinds of benefits the system could realise (BC3), to the patient, these could include improvements to physical well-being from targeted and effective interventions, and to the clinician in-charge the benefit could be the psychological benefit of well-supported decision-making as well as a reduced burden when many patients need to be attended to simultaneously. Indirectly, other patients on the ward and other clinicians may indirectly benefit from this reduced burden, too. The benefits to the hospital or NHS trust could be greater efficiency, and the indirect legal and financial benefits of increased patient safety and a better-supported clinical staff. The benefits to the developer or manufacturer would likely be financial and reputational (which in turn, may indirectly also be a wider societal or economic benefit).

The point of the Beneficence Assurance Argument is that what benefits the use of the system realises, and for which groups of beneficiaries, should be made explicit. The argument strategy (BA1) is that, if the benefits are realised (or are justifiably expected to be realised) for the specified kinds of benefit and beneficiaries, then confidence in the argument’s top-level claim (BG1) of beneficial use of the system will be justified. The purpose of this argument is not to carry out trade-off arguments, e.g. between benefits and the risks that the systems also impose, but to capture the realised (or justifiably expected) benefits of the system as a precursor to those arguments within the Justice Assurance Argument.

The idea is that the context at (BC4) will be communicable as a benefit matrix that feeds into the Justice Assurance Argument. As a first pass, we think that the combinations in the matrix - which articulates which benefits are realised for which beneficiaries - should be supported by justified confidence in the truth of the following claims. First, that the benefit itself has been appropriately described (BG3) - both what kind of benefit it is (BG7) and what the scale of the benefit is, i.e. its weight, significance, or reach (BG8). Second, that the methods used to realise it are themselves appropriate, i.e. that the specific ways in which welfare is advanced by use of the system is appropriate given its purpose or intended context (BG4). Third, that the benefit is of sufficient scale (given the purpose, intended context, and the kind of benefit it is) to warrant inclusion within the benefit matrix as a part of the justification to use the system (BG5). And fourth, that this benefit can be monitored throughout use (BG6).
These are difficult and non-trivial questions that warrant further thought and validation as we develop the model.

5.3. The Non-maleficence Assurance Argument

The ethical principle of non-maleficence imposes an obligation not to inflict unnecessary harm on others (Beauchamp and Childress 2009: 149). This is the traditional arena of safety assurance. Most commonly, safety assurance takes the form of hazard-driven safety processes. A ‘hazard’ is a foreseeable potential harm, or a harm ‘waiting to happen’. In hazard-driven safety processes, engineers derive safety requirements as a means for constraining the design and use of the system so that the risk of hazardous behaviours is reduced to acceptable levels. Most influential safety standards are centred on, and commence with, hazard identification and risk assessment activities (ISO 2011; ISO 2019). As a result, the primary assurance argument in a safety case which has been produced in compliance with these standards tends to be risk-based and centred on the evidence recorded in a hazard log (or similar mechanism), which identifies how the risks associated with each hazard are controlled.

By including safety as a value to be upheld under the principle of non-maleficence, something becomes clear that the engineering community often fails to recognise: safety is not conceptually distinct from ethics and ethics is not therefore something that is ‘added on’ to safety; rather, safety is itself an ethical concern. Safety assurance has historically focused on the risk of physical harm: the obligation to avoid fatalities and bodily injury. But the real-world deployment of AI/AS to date has made it clear that the range of harm to humans from these systems now extends beyond the purely physical. Indeed, arguably the whole field of AI/AS ethics emerged precisely because of the extended range of harms that the systems may cause (Leslie 2019: 4). The Non-maleficence Assurance Argument, which concerns the avoidance of unjustified harm from the use of the system, is given in Figure 7.

Let us now describe the argument in more detail. The top-level claim at (NG1) is that the use of the system does not cause unjustified harm to identified groups of risk-exposed individuals. By ‘risk-exposed’, those parties who will be identified in the contextual artefact (NC2), we mean those whose welfare may be negatively impacted by deployment of the system. Returning to the example of the robo-taxi, at the individual level, the likely risk-exposed people will be: i) end users of the service; ii) other individuals in the ODD; and also iii) human taxi drivers who once picked up fares at the taxi rank at the station and have been replaced. At the group or societal level, they include: iv) the service providers or operators; v) the municipal or city council; vi) developers and manufacturers; and also, indirectly, vii) regulatory bodies and certification agents, which would suffer reputational damage from the deployment of approved but ethically unacceptable systems.

---

19 To be clear, the thought here is that safety is a subset of ethics, not that safety and ethics are equivalent, nor that all ethical considerations are also safety considerations.

20 The opprobrium heaped on the FAA following the Boeing 737 Max accidents illustrates this risk-exposure
Naturally, physical harm is one of the main kinds of possible harm from the use of a robo-taxi, both for users and other individuals in the ODD. But for AI/AS, it is clear that the risk of harm extends beyond the purely physical. The extended range of harm that might be incurred from the use of the given system is contained within the contextual artefact (NC3). Physical harms from safety-critical systems are reasonably well understood already. The introduction of psychological harms, whether some form of trauma (e.g. PTSD) from injury, or addictive behaviours, or stress and anxiety (BS 8611: 2016) is relatively new to the traditional system safety world. And beyond the physical and psychological, other harms should be included within this contextual artefact - these may be invasions of privacy, the harms that arise from
discrimination or algorithmically-exacerbated bias against demographic groups, economic harms, societal harms (e.g. to infrastructure or social cohesion), and environmental harms.

We can now consider what kinds of harms are applicable to our two examples. In the hypothetical robo-taxi case, in addition to the clear physical risks consequential upon unsafe design or operation of the vehicles, there are the possible psychological harms of anxiety or stress to the user if the vehicle behaves erratically or dangerously. The economic exclusion of the replaced human taxi drivers is a serious issue that would need to be captured in the Non-maleficence Assurance Argument; this would also be expected to incur psychological harm to those individuals. Moreover, aggregated over all replaced taxi drivers, use of this AI/AS could become a source of societal harm and economic exclusion, because it removes the need for an established workforce and their source of income. Another source of harm is discriminatory bias. In this context, the bias may arise because the systems have not been adequately trained on representative samples of the population demographic within the ODD. To take just one example, research has shown that the ML-models in self-driving cars have yet to deal with wheelchairs falling over backwards (Financial Times 2020). In addition, questions of indirect discrimination may arise if the use of the systems means that, for example, pollution or traffic congestion is pushed to a less affluent area just outside the ODD. A further environmental hazard is that training ML models has a significant carbon footprint; it is unclear how this will ultimately affect greenhouse gas emissions when balanced against the energy savings such systems may also bring further down the line of deployment (Kaaack et al. 2021). The risk to group actors listed above, such as the developer, service provider, local council, and relevant regulatory body will be largely legal, financial, and reputational.

Looking at the example of AI Clinician, at the individual level, the risk-exposed, identified by the contextual artefact (NC2) would be: i) the patients; ii) the individual clinician in-charge; iii) other patients and clinicians in the operating environment; and at the group level, they would be: iv) the hospital or NHS trust; v) the developer or manufacturer of the system; and vi) the

---

21 Discrimination is typically considered to be an injustice (since its wrong-making feature is generally explained in terms of fairness rather than harm; and not all discrimination necessarily cause harm) - a point we are grateful to Dr Chris Jay for emphasising. But we include discrimination in the Non-maleficence Assurance Argument for two reasons. First, for pragmatic and organisational reasons - because it is a risk that, we believe, would be more practicably dealt with at this stage in this process, and then exported up to the Justice Assurance Argument, rather than considered there independently in the first instance. Second, because while discrimination or bias might not always cause harm, such that a simple assimilation of harm and discrimination would be unsound, it will plausibly always be a hazard or possible source of harm - which is what the Non-Maleficence Assurance Argument is concerned with. Indeed, something this concern prompts for future development of this work is how to ensure that discrimination is also considered with the Justice Assurance Argument.

22 As is well-documented, the problem of algorithmic bias is endemic (see Buolamwini and Gebru 2018; Noble 2018; Eubanks 2018; Wachter-Boettcher 2017; O’Neill 2016). In terms of technical mitigations, measures which mitigate bias in one respect may exacerbate bias in another. Many technical fairness measures are pairwise incompatible (Friedler et al. 2016; Chouldechova 2017). Given that each measure chosen will have significant normative consequences, hard choices will need to be made before any are implemented. Further, the technical community cannot proceed in these endeavours alone; it requires a more public debate and the guidance of regulators, lawyers, philosophers, ethicists, and social scientists (Zimmermann, di Rosa and Kim 2020; Wong 2020; Binns 2018). Best practice and regulatory guidance have not yet been established.
relevant regulatory body or bodies. Turning to the contextual artefact denoted by (NC3), for the patient, the principal risk is physical harm, which, given the criticality of the context, includes loss of life. Physical risk may be compounded by the wrong of discriminatory bias if, for example, the system has not been built or trained in such a way that it is sensitive to variations in symptoms, complications, or accompanying conditions amongst demographic or ethnic groups. The risk to the clinician is psychological, insofar as maintaining appropriate oversight of the system in addition to managing a busy and critical workload may be excessively stressful. The clinician may also bear serious concerns about accountability and liability if they implement, in good faith, a system recommendation that turns out to be wrong. The risk to the hospital or NHS trust, as well as to the developer, manufacturer and regulator, will likely be, as above, legal, financial, and reputational - if, for example, several cases of systems making wrong recommendations lead to litigation, mistrust, and public condemnation.

Thus, the Non-maleficence Assurance Argument becomes an extended safety assurance argument.24 The argument strategy, given at (NA1), is that, if the hazards as sources of harm are addressed, then we can have confidence in the Non-maleficence Assurance Argument's top-level claim that the system does not, or will not, cause unjustified harm (NG1) - by which we mean harm that is not reasonable given the purpose and use of the system, will be warranted.25 As with the Beneficence Assurance Argument, the purpose of this argument is not to carry out trade-off arguments, but to capture what the residual risks of the system are, for which risk-bearers, as a precursor to those arguments within the Justice Assurance Argument.

The idea here, much like with the previous Beneficence Assurance Argument, is that the context element (NC4) will reference a residual risk matrix that feeds into the Justice Assurance Argument. As a first pass, we think this matrix - which articulates which residual risks that remain and for which risk-bearers - should provide the scope for the following claims. First, that the hazards have been appropriately described (NC3) - both the kind of hazard (NC7) and its scale (NC8). Second, that the methods used to mitigate the hazards are themselves appropriate, i.e. that the specific ways in which risk is managed is appropriate given its purpose or intended context (NC3). Third, that these mitigations are sufficient (NC5). And fourth, that the hazards can be monitored throughout use (NC6).

---

23 The implementation and failings of the Horizon IT system at the Post Office, and the subsequent prosecution of over 700 sub-postmasters, shows that such a concern would not be misplaced.

24 Anecdotally, we have noticed that safety engineers are generally comfortable including psychological harm within their sphere of activity, but less so about the other forms of harm. Not only may the structure of the overall Principle-based Ethical Assurance Argument pattern allow different groups to work on different sub-trees of the argument, but different specialists could also work on different elements of the sub-trees concurrently - allowing privacy or algorithmic bias specialists, for example, rather than safety specialists to work on the relevant elements of the Non-maleficence Assurance Argument (whilst still following the same overall methodology). With appropriate communication and transparency, the argument pattern could play a useful role in facilitating and harmonising this multidisciplinary and interdisciplinary approach.

25 This is not to say, however, that the bulk of the overall argument's justification takes place within the Non-maleficence Assurance Argument. Rather, this is a site of partial justification. Residual risks of harm, contained with the residual risk matrix, feed into the Justice Assurance Argument where the argument's final justification of the decision to deploy or use a system takes place.
As with the previous sub-argument, these are difficult and non-trivial questions - particularly given the extended range of harms that use of the systems may incur. They warrant further thought as we develop the model. But it is our working hypothesis that all forms of harm, including discriminatory bias, are amenable to a broadly similar approach - adopted from traditional safety methods - of identifying and describing their sources (hazards), and then implementing appropriate measures to mitigate and manage them. Mitigation of hazards may not always be possible within the Non-maleficence Assurance Argument, however. Indeed the risk matrix is a residual risk matrix. The intent is that these residual risks are set within the wider context of benefits and impacts on personal autonomy in the Justice Assurance Argument, and reasoned about and evaluated there.

5.4. The Personal Autonomy Assurance argument

The meaning of ‘autonomy’ in moral and political philosophy is almost too rich to admit of easy characterisation, but virtually all theories highlight two essential conditions: freedom and the capacity for intentional action (Beauchamp and Childress 2009: 100). A person’s autonomy involves the freedom and capacity of human beings to direct their own lives, to exercise their deliberative capacities, to live according to their own reasons and motives, and to be responsible for their decisions and actions. Personal autonomy is therefore central to the moral agency of human beings. Within medical ethics, respect for autonomy is generally taken to be “the first principle amongst equals” (Gillon 2003) and it underscores the emphasis on a patient’s ‘informed consent’ with respect to treatment or intervention. Determining the ethical acceptability of a system cannot simply consider risk and benefit, but should also consider what respect is given to the personal autonomy of those affected by the use of the system. The Personal Autonomy Assurance argument is presented in Figure 8.

We can describe how this argument works in more detail - but ask readers to note also that the Personal Autonomy Assurance Argument also still requires further development. The top-level claim of the argument is that the use of the system does not unduly constrain the personal autonomy of identified groups of ‘autonomy risk-exposed’ individuals (AG1). By ‘autonomy risk-exposed’ (AC2), we mean a subset of the risk-exposed described in the Non-maleficence Assurance Argument. These are the individuals whose autonomy may be constrained by the operation of the system. In the example of the robo-taxi, the ‘autonomy risk exposed’ would be the individual users and other individuals in the ODD; in the case of the AI Clinician, they would be the patients and the individual clinicians, and, possibly, others individuals on the ward (imagine another clinician, for example, who cannot avoid the intrusive presence of the system when treating other patients). The types of constraint on personal autonomy that we consider (AC3) include constraints on personal choice, informed consent, and meaningful control.
Figure 8 Personal Autonomy Assurance Argument

The argument strategy of the Personal Autonomy Assurance Argument (AA1) is that confidence in the top-level claim that the use of the system does not unduly constrain people's personal autonomy will be justified when challenges to personal autonomy have been managed. Any challenges that cannot be managed should also be explicitly identified. As before, the purpose of this argument is not to carry out trade-off arguments, e.g. between constraints on personal autonomy and other risks or benefits. The purpose of the Personal Autonomy Assurance Argument is to capture the constraints on personal autonomy that have been managed, and those that remain, as a precursor to those arguments within the Justice Assurance Argument. As with the Beneficence and the Non-maleficence Assurance Arguments, what we would like to achieve here is a personal autonomy matrix, which gets exported up to the Justice Assurance Argument. The context element (AC4) will reference a personal autonomy matrix that feeds into the Justice Assurance Argument.

We think this matrix - which articulates the managed and residual constraints on personal autonomy, and for which individuals - should provide the scope for the following claims. First, that the personal choice of the relevant individuals has been considered and catered for, as appropriate (AG3). For example, in the robo-taxi example, the ODD will be deliberately scoped and 'enveloped' to facilitate optimal system performance (Floridi 2019), but the risk of this may be reduced freedom for those who live or work within that environment; their choice in the matter is something that should be explicitly considered. In the AI Clinician case, there are questions, for example, around what meaningful choice a clinician has with respect to
using the system. Second, that the relevant individuals’ capacity to give informed consent has likewise been considered and addressed (AG4). Sometimes, a kind of hypothetical informed consent will be given by proxies such as regulators. Generally, this occurs for very complex systems that have the potential for very widespread harm, e.g. aircraft, and the regulators can be seen as acting on behalf of the public in approving (or withholding approval for) the use and outcomes associated with the AI/AS. In the case of the robo-taxi example, aside from the ability of individual users to decide whether or not to use the service on a given occasion, informed consent would likely be of this kind. In the case of the AI Clinician, the patient may be too ill to give consent; but in such cases, one would expect that the normal assessment of capacity and ‘best interests’ decision-making would apply.\footnote{‘Best interests’ is detailed in the Mental Capacity Act 2005, Part 1, Section 4. It concerns decision-making about life-sustaining treatment when a person is not able to participate in that decision themselves, and includes considerations of what their wishes reasonably might have been, as well the views of anyone named by them as someone to be consulted, or of anyone engaged in caring for them.}\footnote{Conversely, ‘too much’ control here might mean the AI/AS ends up being redundant (with thanks to Dr Tom Lawton for this point).} Informed consent can also be more broadly construed to ensure that the system does not unduly nudge or coerce people, including expert users, into decisions (Burr, Christiani and Ladyman 2018). This is a form of harm that was excluded from the Non-maleficence Assurance Argument and sits naturally within the Personal Autonomy Assurance Argument.

One of the recurrent themes from the many sets of ethical principles published for AI/AS is that there should be human control of the technology (Fjeld et al. 2020). This is captured in the claim expressed in the third sub-goal at (AG5). By ‘meaningful control’ we mean two things. First, that users can intervene in a system that is going awry - they can either stop it or change its course - and that this is not just a literal or theoretical capacity but one that they can exercise effectively in practice (AG6) Second, that users can ensure, in actionable ways, that the system is acting as intended and desired (AG7). Drawing on philosophical work in the domain of moral responsibility (which requires meaningful control over actions and events), we might call these ‘regulative’ and ‘guidance’ control, respectively (Santoni de Sio and van den Hoven 2018; Fischer and Ravizza 1998).

In the running example of the robo-taxi, considerations at (AG6) would include whether an individual user of the service could stop the vehicle if needed or desired, and at (AG7) whether individual operators in remote control centres, who have oversight of the whole fleet, can adequately ensure that the systems are acting as intended, and make this oversight actionable. In the example of the AI Clinician, the main consideration at (AG6) would be whether the clinician has the capacity and requisite knowledge to stop the system (more correctly, to discount its outputs since it is a decision-support system). At (AG7) considerations might include questions such as whether the clinician can influence how changes in the inputs would propagate through to output changes,\footnote{Conversely, ‘too much’ control here might mean the AI/AS ends up being redundant (with thanks to Dr Tom Lawton for this point).} or whether the clinician can monitor what features of the patient’s state the system is responding to when making recommendations.
It should be noted, however, that this emphasis on meaningful human control could be misused if it is taken to justify using humans-in-the-loop as a ‘moral crumple zone’ or ‘liability sink’, whereby moral blame or legal liability is unfairly attributed to and absorbed by the nearest human operator when a system malfunctions (Elish 2019). Though meaningful human control should increase these individuals’ capacity to intervene effectively, it would still be unjust to hold that they (and they alone) would always be culpable for the consequences of failing to do so. The risk of this can already be seen in cases involving AVs. Take, for example, the first fatality of a pedestrian caused by a self-driving car: Elaine Herzberg in Tempe, Arizona, in March 2018. While Arizona prosecutors decided that Uber, the developer and operator of the vehicle, would not be criminally liable for the death, the safety driver, Rafaela Vasquez, has had charges of negligent homicide brought against her (BBC 2020; Wamsley 2019), which might be viewed as an example of someone becoming a ‘liability sink’. The most recent report by the Law Commission has sought to bring in clarity about transitions between autonomy and human driving, and this can be seen as a response to the problems of the ‘moral crumple zone’ and the ‘liability sink’ (Law Commission 2022).

6. Transparency

Perhaps a controversial aspect of our proposal is that transparency is not included as a fifth principle in its own right. Our reason for this is that we view transparency ultimately as a means to an end rather than an end in itself. Whereas justice, beneficence, non-maleficence, and respect for personal autonomy are intrinsically important, transparency is instrumentally important. As we discussed in section 4, the fact that the wrong kind of transparency, or an inappropriate degree of transparency, can sometimes work against ethical concerns makes it reasonable to incorporate it within the overall argument as a carefully targeted value that enables the other principles to be enacted and realised. How this should be done, and how the goal of appropriate system transparency should be decomposed within this argument module, will be a research question for us as we continue to develop the model.

We use ‘transparency’ to refer to the overall state of visibility around and about the system. This includes visibility of human decision-making (which might be called ‘governance transparency’) and visibility of what is going on ‘under the hood’ of the AI/AS (which might be called ‘system transparency’). This aligns with the definition used within the recently published IEEE P7001 standard on transparency, as “the transfer of information from an autonomous system or its designers to a stakeholder, which is honest, contains information relevant to the causes of some action, decision or behaviour and is presented at a level of abstraction and in a form meaningful to the stakeholder” (IEEE P7001: 14). By contrast, Arrieta and colleagues define transparency as an understandable model of the system (Arrieta et al. 2019: 5). We think there would be significant value to the research community in standardisation of the meaning given to the many apparently synonymous terms that are used for system transparency: explainability, interpretability, intelligibility, comprehensibility,

---

28 The term ‘liability sink’ comes from Dr Tom Lawton.
29 With thanks to Ivan Kyambadde for bringing this to our attention.
and, sometimes, accountability (in the sense of ‘giving an account of’ rather than in the sense of being held morally or legally accountable). Koshiyama and colleagues make a distinction between ‘explainability’ - the extent to which the system can be understood in human terms - and ‘interpretability’ - the extent to which the causal and logical process of the system can be observed in order to glean how the model works (Koshiyama et al. 2021: 11). Interpretability enables forward-looking predictions to be made for a given set of input or model parameters; it can also allow for backward-looking descriptions of how a specific output was reached.

Different types of governance transparency and system transparency will be required by different stakeholders, at different points in the lifecycle, for different reasons (McDermid et al. 2021; IEEE P7001; Ward and Habli 2020; ICO-Turing 2020). Though we do not, at this stage, open up the transparency argument module - this is a matter for future, carefully considered research - the following generalisations can be made. First, the whole exercise of an ethical assurance argument is an exercise in transparency, since there is both an explicit representation of the logical flow of the argument and its rationale, and a visibility of the facts, data, and evidence that have been used to support justification in its claims. Second, and more precisely, there will often be different transparency requirements (both transparency of the human decision-making processes, and transparency as a property of the AI/AS itself, or ‘AI explainability’) for each of the four principle-based arguments. For the Justice Assurance Argument, traceability and documentation of - indeed, honesty about - the human decision-making process will be key. Human decision-makers should also be expected to explain why they chose to design and develop the system as they did, and why they consider it ethically defensible to put forward the system for use in the intended context. For the Beneficence Assurance Argument, visibility on such things as ways in which operators can ensure that the system as used realises its benefits will be required. For the Non-maleficence Assurance Argument, explanations will likely be the most detailed, including, for example, the kind of explanations that support determinations of whether a system embodies or may exacerbate discriminatory bias. For the Personal Autonomy Assurance Argument, explanations will need to be of the right kind to inform consent, where appropriate, and to enable meaningful control (i.e. present the most useful and relevant information, at the right level of abstraction for the user). There must also be an intelligible basis upon which an affected party could contest or challenge an AI/AS output or outcome. And recent emphasis on ‘explainable AI’ should not occlude the fact that all of these computational techniques to explain or interpret the systems sit within a wider framework of human accountability (McDermid et al. 2021).

7. Conclusion

In this paper, we have combined 1) an analysis of how the four principles of biomedical ethics can be made relevant to and can be adapted to the ethics of AI/AS with 2) an ethical assurance argument for the use of an AI/AS. The result is a prototype of a Principle-based Ethical Assurance Argument pattern. This, we think, is a promising way to try to bridge the gap between principles and practice.
While this has been an exercise in substance - identifying the content of the ethical requirements - it has also been an exercise in organisation - establishing how a diverse range of ethical concerns can be effectively and coherently addressed in one piece. The ambition is that this will help designers, engineers, developers, and regulators reason about the complex question of ethical acceptability as a whole, and take concrete, ‘joined up’ steps in that direction. In being explicit, dialogical, and reviewable, this methodology also carries the benefit of being able to assure an AI/AS through-life. In addition, this principle-based methodology looks promising for meeting new regulations on the horizon. One dimension of ongoing research is mapping the model against requirements in draft legislation, such as the European Commission’s proposed Artificial Intelligence Act, as well as the forthcoming white paper setting out the UK’s position on AI regulation.

The central ethical question about these new technologies is when the decision to deploy an AI/AS would be justified. Given the novelty of the systems and the conditions of uncertainty in which, and for which, they are deployed, it is highly unlikely that regulators will get the answer to this question right with foresight alone.\textsuperscript{30} The most practical way forward is to deploy ethically defensible systems and learn from experience during their operation.\textsuperscript{31} We believe that the Principle-based Ethical Assurance Argument pattern presented in this paper could provide precisely the sort of methodology required for informing and taking that crucial step - and, because of the amenability of the method to reviewability, thus for incorporating learning from that initial deployment.

As we continue to develop the model, our focus will be on the crucial Justice Assurance Argument and the Personal Autonomy Assurance Argument, as well as the Transparency Argument module. In addition, we are working to develop, improve, and validate the model as whole by applying it to case studies of AI/AS across a range of domains of application. This will help us further to test its promise as a practical methodology, and improve it as necessary, to achieve ethically acceptable AI and autonomous systems.

\textsuperscript{30} This was a theme that emerged during a workshop run by the Johns Hopkins’ Institute for Assured Autonomy in February 2022, attended by two of the authors.

\textsuperscript{31} Through-life assurance is a standard principle in safety-critical industries although the practice can be somewhat lacking. However, this principle is emphasised further for AI/AS given the high level of uncertainty about the actual performance of these systems in complex and often open operating environments - and we see the issue of through-life monitoring and management of AI/AS being a critical research direction.
References

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

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58: 82-115

BBC, 2020. Uber's self-driving operator charged over fatal crash, 16 September 2020. Retrieved from: https://www.bbc.co.uk/news/technology-54175359

Beauchamp, T. and Childress, J. 1979. Principles of biomedical ethics (1st ed.) New York, NY: Oxford University Press

Beauchamp, T. and Childress, J. 2009. Principles of biomedical ethics (6th ed.). New York, NY: Oxford University Press

Binns, R., 2018. Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency (pp. 149-159).

BS 8611:2016 Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems (British Standard)

Buolamwini, J. and Gebru, T., 2018, January. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings in Machine Learning Research 81: 1-5

Burr, C., Cristianini, N. and Ladyman, J., 2018. An analysis of the interaction between intelligent software agents and human users. Minds and Machines, 28(4): 735-774

Burr, C. and Leslie, D., 2021. Ethical Assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies. (preprint arXiv:2110.05164)

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

Centre for Data Ethics and Innovation. 2021. The roadmap to an effective AI assurance ecosystem https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem
Choi, B. and Pak, A. 2006. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. Clinical and Investigative Medicine, 29(6):351-64

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2): 153-163

Crisp, R. 2021. Well-Being, The Stanford Encyclopedia of Philosophy (Winter 2021 Edition), Edward N. Zalta (ed.), (https://plato.stanford.edu/archives/win2021/entries/well-being/) <accessed 30 November 2021>

Daniels, N., 1979, ‘Wide reflective equilibrium and theory acceptance in ethics’, Journal of Philosophy, 76(5): 256–82

Data Protection Act 2018 (UK) https://www.gov.uk/government/collections/data-protection-act-2018

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

Engineering Council and Royal Academy of Engineering 2017. Statement of ethical principles https://www.engc.org.uk/media/2334/ethical-statement-2017.pdf

Eubanks, V. 2018. Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press

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/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

European Commission 2021. Annexes: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF

Floridi, L. 2019. What the near future of Artificial Intelligence could be. Philosophy & Technology, 32(1): 1-15
Floridi L. and Cowls J. 2021. A unified framework of five principles for AI in society. In: Floridi L. (eds) Ethics, Governance, and Policies in Artificial Intelligence. Philosophical Studies Series, vol 144. Springer

Floridi, L., Cowls, J., Beltrametti, M. Chatila, R., Chazerand, P. Dignum, V. Luetge, C., Madelin, R., Pagallo, U. Rossi, F. and Schafer, B. 2018. AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines 28(4): 689–707

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M., 2020. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, (2020-1)

Friedler, S.A., Scheidegger, C. and Venkatasubramanian, S. 2016. On the (im) possibility of fairness. (arXiv:1609.07236)

General Data Protection Regulation 2016 (GDPR) (EU) https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1374-1-1

Gillon, R. 1994. Medical ethics: four principles plus attention to scope. British Medical Journal 309: 184

Gillon, R. 2003. Ethics needs principles—four can encompass the rest—and respect for autonomy should be “first among equals”. Journal of Medical Ethics 29(5): 307-312

Assurance Case Working Group, 2021. Goal structuring notation community standard (version 2). https://scsc.uk/r141C:1

Goodenough, J., Weinstock, C. and Klein, A. 2012. Toward a theory of assurance case confidence. (September 2012) Technical Report CMU/SEI-2012-TR-002, Software Engineering Institute, Carnegie-Mellon University

Habli, I., Alexander, R. and Hawkins, R. 2021. Safety cases: an impending crisis? In Safety-Critical Systems Symposium (SSS’21). York, 2021

Hansson, S.O., 2003. Ethical criteria of risk acceptance. Erkenntnis, 59(3): 291-309

Hansson, S.O., 2018. How to perform an ethical risk analysis (eRA). Risk Analysis, 38(9): 1820-1829

Hauer, M.P., Adler, R. and Zweig, K. 2021. Assuring fairness of algorithmic decision making. 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2021: 110-113
Hawkins, R., Kelly, T., Knight, J. and Graydon, P. 2011. A new approach to creating clear safety arguments. In *Advances in Systems Safety* (pp. 3-23). Springer, London

Health and Safety at Work etc Act 1974 (UK)  
https://www.legislation.gov.uk/ukpga/1974/37/enacted

HM Government. 2021. UK National AI Strategy  
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf

Holland, S. 2015. *Public health ethics*. John Wiley & Sons.

House of Lords. 2018. AI in the UK: ready, willing and able? Technical Report of the House of Lords Select Committee on Artificial Intelligence, UK  
https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

ICO/Turing. 2020. Explaining decisions made with AI.  
https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/

ISO. 2011. International Standard 26262 Road Vehicles—Functional Safety

ISO. 2019. International Standard 14971:2019 Medical devices — Application of risk management to medical devices

Jobin, A., Ienca, M. and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), pp.389-399

Kaack, L., Donti, P., Strubell, E., Kamiya, G., Creutzig, F. and Rolnick, D. 2021. Aligning artificial intelligence with climate change mitigation.  
https://hal.archives-ouvertes.fr/hal-03368037

Kelly, T. 1998. Arguing Safety — A Systematic Approach to Safety Case Development", Doctoral Thesis, Department of Computer Science, University of York

Kelly, T. and McDermid, J. 1997. Safety Case Construction and Reuse Using Patterns. In: Daniel P. (ed) *Safe Comp 97*. Springer, London.

Komorowski, M., Celi, L.A., Badawi. O., Gordon, A.C., Faisal, A. A. 2018. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*. 2018. 24(11):1716–20
Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavely, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S. and Lomas, E. 2021. Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. http://dx.doi.org/10.2139/ssrn.3778998

Law Commission. 2022. Automated vehicles: joint report (Law Commission of England and Wales and Scottish Law Commission) https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2022/01/Automated-vehicles-joint-report-cvr-03-02-22.pdf

Leslie, D. 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

McDermid, J.A., Jia, Y., Porter, Z. and Habli, I. 2021. AI explainability: the technical and ethical dimensions. Philosophical Transactions A: Mathematical, Physical and Engineering Sciences

McDermid, J.A., Porter, Z. & Jia, Y. 2022. Consumerism, Contradictions, Counterfactuals: Shaping the Evolution of Safety Engineering. In M Parsons & M Nicholson (eds), Safer Systems: The Next 30 Years: Proceedings of the 30th Safety-Critical Systems Symposium. Safety Critical Systems Club, 170: 15-36

McDermid, J.A. 1994. Support for safety cases and safety arguments using SAM. Reliability Engineering & System Safety, 43(2): 111-127

Menon, C. and Alexander, R. 2020. A safety-case approach to the ethics of autonomous vehicles. In Safety and Reliability 39(1): 33-58

Mental Capacity Act 2005 (UK) https://www.legislation.gov.uk/ukpga/2005/9/contents

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, November 2019: 501-507

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A., 2019. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics (published online 11 December 2019)

Noble, S.U. 2018. Algorithms of oppression. New York University Press
O’Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group

Rawls, J., 1951. Outline of a decision procedure for ethics. *The Philosophical Review*, 60(2): 177-197

Rawls, J. 1971. *A theory of justice*. Harvard University Press

Santoni de Sio, F. and Van den Hoven, J., 2018. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI* (5)

Savage, N., 2020. The race to the top among the world’s leaders in artificial intelligence. *Nature*, 588 [https://www.nature.com/articles/d41586-020-03409-8](https://www.nature.com/articles/d41586-020-03409-8)

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

The Health Foundation. 2012. Using safety cases in industry and healthcare [https://www.health.org.uk/publications/using-safety-cases-in-industry-and-healthcare](https://www.health.org.uk/publications/using-safety-cases-in-industry-and-healthcare)

Toulmin, S. 2003. *The uses of argument*. Cambridge University Press

Turilli, M. and Floridi, L. 2009. The ethics of information transparency. *Ethics and Information Technology* 11: 105–112

Tricot, R. 2021, Venture capital investments in artificial intelligence: Analysing trends in VC in AI companies from 2012 through 2020, OECD Digital Economy Papers, No. 319, OECD Publishing, Paris, [https://doi.org/10.1787/f97beae7-en](https://doi.org/10.1787/f97beae7-en).

United Nations. 2021. United Nations Activities of Artificial Intelligence (AI) 2021 Report. [https://www.itu.int/hub/publication/s-gen-unact-2021/](https://www.itu.int/hub/publication/s-gen-unact-2021/)

Wachter-Boettcher, S. 2017. *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech*. WW Norton & Company

Wamsley, L. 2019. Uber not criminally liable in death of woman hit by self-driving car, prosecutor says. 6 March 2019. NPR. Retrieved from: [https://www.npr.org/2019/03/06/700801945/uber-not-criminallyliable-in-death-of-woman-hit-by-self-driving-car-says-prosec](https://www.npr.org/2019/03/06/700801945/uber-not-criminallyliable-in-death-of-woman-hit-by-self-driving-car-says-prosec)

West, D.M. and Allen, J.R., 2018. How artificial intelligence is transforming the world. *Brookings Institution Report*
Wong, P.H., 2020. Democratizing algorithmic fairness. *Philosophy & Technology*. 33(2): 225-244

Zimmerman, A., Di Rosa, E. and Kim, H., 2020. Technology can’t fix algorithmic injustice. *Boston Review*. (January 9, 2020)