AI & SOCIETY (2025) 40:185–198
https://doi.org/10.1007/s00146-024-01866-7

OPEN FORUM

Ethical governance of artificial intelligence for defence: normative tradeoffs for principle to practice guidance

Alexander Blanchard[1] · **Christopher Thomas**[2] · **Mariarosaria Taddeo**[1]

Received: 22 July 2023 / Accepted: 5 January 2024 / Published online: 19 February 2024

Abstract
The rapid diffusion of artificial intelligence (AI) technologies in the defence domain raises challenges for the ethical governance of these systems. A recent shift from the *what* to the *how* of AI ethics sees a nascent body of literature published by defence organisations focussed on guidance to implement AI ethics principles. These efforts have neglected a crucial intermediate step between principles and guidance concerning the elicitation of ethical requirements for specifying the guidance. In this article, we outline the key normative choices and corresponding tradeoffs that are involved in specifying guidance for the implementation of AI ethics principles in the defence domain. These correspond to: the AI lifecycle model used; the scope of stakeholder involvement; the accountability goals chosen; the choice of auditing requirements; and the choice of mechanisms for transparency and traceability. We provide initial recommendations for navigating these tradeoffs and highlight the importance of a pro-ethical institutional culture.

Keywords AI ethics · AI lifecycle · AI in defence · Principle to practices · AI governance · Guidelines

1 Introduction

The diffusion of artificial intelligence (AI) technologies in the defence domain raises questions and challenges for the ethical governance of these systems. As in other domains (Jobin et al. 2019), initiatives have multiplied to address these challenges, including the publication of sets of ethics principles. This in turn has motivated a shift from the *what* to the *how* of AI ethics (Floridi 2019, 185) and, increasingly, the development of guidelines and tools for compliance with those principles.

However, the focus on guidelines and tools in the shift from *what* to *how* leaves unaddressed crucial intermediate questions for developing sound and effective guidelines: it has overlooked the profound normative issues that are involved in setting out guidelines for the implementation of AI ethics principles. This stems from the high-level and foundational nature of AI ethics principles (Morley et al. 2021), meaning that they can be interpreted following different methodologies, in a way that entails different (and difficult) normative tradeoffs (Taddeo et al. 2024). For example, when applied to specific cases AI ethics principles may generate tensions requiring balancing of the principles, with the desirable balance impossible to identify through recourse to the principles or tools alone (Whittlestone et al. 2019). At a more granular level, varying the metrics used to measure compliance with AI ethics principles will also vary the normative outcomes. The question emerges, then, as to what kind of criteria should shape the creation of the guidelines and the choice of tools.

In this article, we outline the normative choices and corresponding tradeoffs that are involved in specifying guidelines for the implementation of AI ethics principles in the defence domain. To do so, we undertake a qualitative systematic review of AI governance literature to identify the key features for the ethical governance of AI that should inform the interpretation and application of principles into practice. We identify five areas where key normative tradeoffs are made: (1) the model of the AI lifecycle; (2) the scope of stakeholder involvement; (3) accountability goals chosen; (4) the choice of auditing requirements; (5) in the choice of mechanisms for transparency and traceability mechanisms. The nature of these requirements and the tradeoffs they entail, as well as preliminary recommendations for navigating/balancing these

✉ Mariarosaria Taddeo
mariarosaria.taddeo@oii.ox.ac.uk

[1] Alan Turing Institute, London, UK

[2] Oxford Internet Institute, University of Oxford, Oxford, UK

tradeoffs, are detailed in Sect. 3. Prior to that, in Sect. 2 we provide an outline of recent efforts in the defence domain to operationalise ethics principles into practice and highlight the limitations of some of these approaches. We conclude our analysis in Sect. 4.

2 From AI ethics principles to practice

In the recent years, efforts to define AI ethics principles have multiplied (Jobin et al. 2019; Floridi and Cowls 2019). The defence domain is no exception, with UK (Ministry of Defence 2022), US (DIB 2020), and Australian (Devitt et al. 2020) national defence institutions, and NATO,[1] issuing their own AI ethics principles. More broadly, AI ethics principles have been criticised for being too abstract to offer concrete guidance to the actual design, development and use of AI systems (Coldicutt and Miller 2019; Peters 2019). Munn (2022) has gone as far as to say that AI ethics principles are "meaningless," "isolated" and "toothless." In the same vein, the efficacy of ethical principles to inform decision-making has been called into question. For example a study including software engineering students and professional software developers showed no statistically significant difference between survey responses from those who read a code of ethics and those who did not (McNamara et al. 2018, 4).

The lack of applicable and effective guidance on how to apply AI ethics principles is particularly problematic in domains like defence and national security, where ethical risks related to the use of AI systems can be severe and may put individual rights and democratic values under sharp devaluative pressure (Taddeo 2013; Taddeo 2015; Blanchard and Taddeo 2023). A lack of guidance for application to concrete cases may also lead to AI ethics principles being seen as

> […] extraneous, as surplus or some kind of "add-on" to technical concerns, as an unbinding framework that is imposed from institutions "outside" of the technical community (Hagendorff 2020, 113).

This, in turn, may reduce AI ethics efforts to a meaningless façade, voided of any concrete outcomes and induce malpractices, like ethics bluewashing[2] (Floridi 2019, 185).

These concerns have motivated a shift from the *what* to the *how* of AI ethics (Georgieva et al. 2022) with a nascent body of literature in the defence domain focussed on developing AI ethics tools and processes to implement AI ethics principles. These include: the US DoD Responsible AI Strategy and Implementation Pathway (2022), US DoD Defence Innovation Unit (DIU) Responsible AI Guidelines in Practice (2021), and Australia's Defence Science and Technology Group 'A Method for Ethical AI in Defence' (Devitt et al. 2020).[3] Here we briefly outline the guidelines and note some initial limitations, before considering in more detail the normative questions entailed in specifying guidelines for operationalising defence AI ethics principles.

2.1 Existing defence AI ethics guidance

The DoD's 'Responsible AI Strategy and Implementation Pathway' is developed in line with the DoD's stated ethical principles: responsible, equitable, traceable, reliable, governable (DoD Responsible AI Working Council 2022). The pathway builds on the DoD's existing infrastructure for technology development and governance, such as sound software engineering practices and robust data management, and reflects an "enterprise-wide approach" prescribing responsibilities for stakeholders across the DoD. The operationalisation of the principles is structured around six tenets with associated goals (Department of Defense 2022, 9–10) (see Table 1):

To implement the six tenets, the DoD prescribe 'Lines of effort' (LOEs), which direct actions to implement best practices and standards, tasking the Department to develop new approaches where necessary. The LOEs are accompanied by overarching goals, the identification of 'Office(s) of Primary Responsibility' (OPRs), i.e., stakeholders responsible for implementation, and estimated timelines for implementation. Figure 1 shows the interaction between the first tenet, the two levels of an LOE, and the corresponding OPR in the RAI implementation pathway. Due to their complex, cross-cutting nature, implementing the LOEs requires department-wide and external stakeholder input.

The Defence Innovation Unit (DIU) (Dunnmon et al. 2021) develops a set of questions for step-by-step guidance for DoD stakeholders, including AI vendors and program managers, to ensure AI programs align with DoD ethical principles, whilst also ensuring that fairness, accountability,

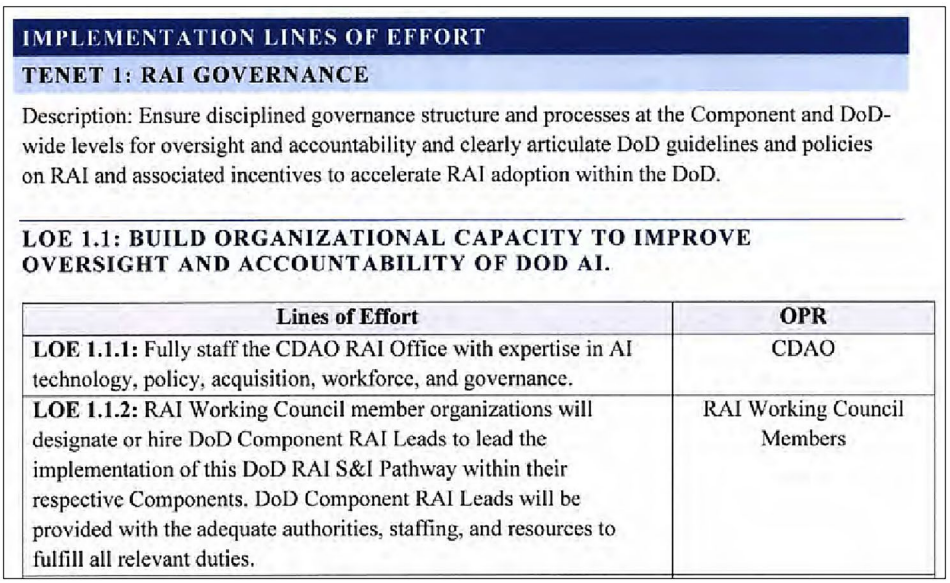
[1] https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html

[2] Defined as, "the malpractice of making unsubstantiated or misleading claims about, or implementing superficial measures in favour of, the ethical values and benefits of digital processes, products, services, or other solutions in order to appear more digitally ethical than one is" (Floridi 2019, 187).

[3] Whilst a number of defence organisations have developed sets of AI ethics principles (see above), to the best of our knowledge these are the only methodologies for applying AI ethics principles provided by defence organisations. See (Canca 2023) for an implementation framework not provided by a defence organisation.

Table 1 The six tenets with associated goals for the applications of the DoD RAI principles ((Department of Defense 2022, 9–10)

	Tenet	Goal
1	Responsible Artificial Intelligence [RAI] governance	Modernising governance structures and processes to allow for context specific, continuous oversight of AI
2	War fighter trust	System operators have a standard level of familiarity with a technology, enabling justified confidence in AI systems and capabilities
3	AI product and acquisition lifecycle	Appropriate care ensures potential risks can be ameliorated and unintended consequences reduced, whilst enabling the development needed to meet the National Defence Strategy
4	Requirements validation	Ensures that AI capabilities are aligned with operational needs whilst addressing relevant risks
5	Responsible AI ecosystem	Promotes a shared understanding of responsible AI, through domestic and international engagement
6	AI workforce	Ensures that DoD AI workforce possess appropriate understanding of the technology and implementing the RAI, commensurate with their duties



IMPLEMENTATION LINES OF EFFORT

TENET 1: RAI GOVERNANCE

Description: Ensure disciplined governance structure and processes at the Component and DoD-wide levels for oversight and accountability and clearly articulate DoD guidelines and policies on RAI and associated incentives to accelerate RAI adoption within the DoD.

LOE 1.1: BUILD ORGANIZATIONAL CAPACITY TO IMPROVE OVERSIGHT AND ACCOUNTABILITY OF DOD AI.

Lines of Effort	OPR
LOE 1.1.1: Fully staff the CDAO RAI Office with expertise in AI technology, policy, acquisition, workforce, and governance.	CDAO
LOE 1.1.2: RAI Working Council member organizations will designate or hire DoD Component RAI Leads to lead the implementation of this DoD RAI S&I Pathway within their respective Components. DoD Component RAI Leads will be provided with the adequate authorities, staffing, and resources to fulfill all relevant duties.	RAI Working Council Members

Fig. 1 Section of the implementation lines of effort for Tenet 1: RAI Governance, (DoD Responsible AI Working Council 2022, 19)

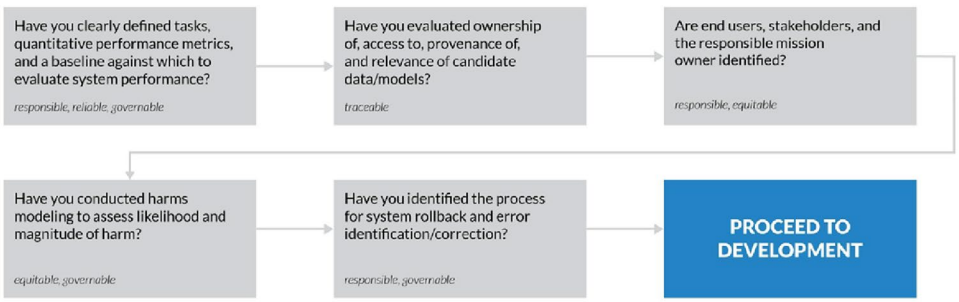


Fig. 2 DIU RAI Guidelines Phase 1 Planning Workflow, (Dunnmon et al. 2021, 8)

and transparency are considered at each step in the development cycle.

The guidelines are developed with a set of 'foundational tenets' to help refine problem formulation and to maximise benefit for national defence whilst aligning with US laws, norms, and values. These are set out in a graphical workflow that visualises specific questions for stakeholders at each lifecycle phase. Each phase is then supplemented with a worksheet that serves as a documentation and verification mechanisms across planning, development, and deployment phases. For example, the planning phase requires personnel from the government agency requesting the system to collaborate with the program manager to define the system's prospective functionality, resources required, and the operational context. This phase has five

key lines of enquiry according to the DoD ethical principles (Dunnmon et al. 2021, 8), as shown in Fig. 2.

Lastly, 'A Method for Ethical AI' proposed by the Australian Defence Science and Technology Group. It aims to ensure accountability for considering ethical risks, assigning specific people to each risk, and making humans accountable for decisions concerning the mitigation of ethical risks. Based on the findings from a stakeholder workshop, the authors propose five 'facets' as broad areas of inquiry—responsibility, governance, trust, law, traceability—with corresponding prompt questions, and then a method comprised of three tools (Table 2).

Given the pace of AI development and ethical risks it poses, initiatives for the implementation of ethics principles are welcomed. Individually, the different guidelines present their own benefits and limitations. The DoD guidelines, for example, delineate the institutional attitude towards the adoption of AI, but they do not offer specific guidance to address the problems that may emerge in applying the principles to specific cases, such as addressing tensions between ethical principles. This means that responsibility for making complex ethical assessments is devolved onto practitioners who may lack the necessary expertise. In addition, mechanisms for stakeholder inclusions are underspecified. This may lead to a negative public perception of the effort to develop responsible use of AI, in turn undermining concrete attempts to develop RAI in defence.

The DIU method aims to address the limitations in the DoD implementation guidance by utilising the Responsible Research Innovation (RRI) framework. This enables the DIU guidelines to enjoy the benefits of the RRI framework—such as the reflective approach which avoids reducing ethical compliance to a box-ticking exercise—but it also inherits two significant pitfalls of the RRI framework (Hajer 2003; Stilgoe et al. 2013). First, the DIU's guidance embeds normative elements in its questions—e.g. "have you *clearly* defined tasks?"—without giving guidance as to how to address those elements. Without such criteria, answers to these questions will remain vague and unsatisfactory. Similarly, a question like "are end users, stakeholders and responsible mission owners identified?" presupposes the specification of the criteria and procedure for identifying stakeholders and their interests which, as detailed in Sect. 3, is far from a trivial task.

Second, the DIU guidance risks shifting the burden for the responsible use of AI from institutions to sole individuals or groups of individuals. This means that, like the DoD guidance, decisions about what is ethically acceptable are left entirely to local decision-makers (researchers, developers, operators, etc.) who may lack the relevant expertise in AI ethics. This approach may work when applied to research and innovation in general (for which the RRI framework was originally conceived), but this is unacceptable when

Table 2 Components of Method for Developing AI Ethically in Defence, (Devitt et al. pp.32–34)

Ethical AI for Defence Checklist for Development of Ethical AI Systems	Ethical AI Risk Matrix	Legal and Ethical Assurance Program Plan (LEAPP)
(a) Describe the military context in which the AI will be employed	Define the activity you are undertaking. Indicate the ethical facet and topic the activity is intended to address. Estimate the risk to the project objectives if issue is not addressed?	Provided by contractors where risk assessment of AI project is above certain threshold
(b) Explain the types of decisions supported by the AI		Data Item Description (DID) provides guidance to contractors to develop ethical assurance programs for complex defence AI systems
(c) Explain how the AI integrates with human operators to ensure effectiveness and ethical decision making in the anticipated context of use and countermeasures to protect against potential misuse	Define specific actions you will undertake to support the activity. Provide a timeline for the activity. Define action and activity outcomes. Identify the responsible party(ies). Provide the status of the activity	DID provides defence organisation with visibility into the legal and ethical planning, supporting risk assessment and providing input into internal planning, including weapons reviews under Article 36 of Additional Protocol 1
(d) Explain frameworks to be used		
(e) Employ subject matter experts to guide AI development		
(f) Employ appropriate verification and validation techniques to reduce risk		

applied to high-risk domains like defence. The RRI framework aims to foster critical thinking and reflection on the potential implications of research and development. It does not aim to offer concrete guidance as to what one ought or ought not to do to mitigate ethical risks of specific cases.

The Australian Defence Science and Technology Group method also presents the first of these limitations, failing to indicate how ethical requirements are to be elicited from the described facets (i.e. principles). Moreover, the wider governance process that ensures the integrity and repeatability of the method are unspecified. Altogether, the method falls short of providing the specific guidance that accountable decision-makers require to mitigate ethical risks and discharge their responsibility.

Some of these limitations may be addressed through revisions of the guidance. For instance, the DIU ought to provide criteria for addressing the normative elements embedded in its guidance questions. This is not a straightforward task, and the oversight itself points to a more fundamental problem that has arisen in the shift from the *what* to the *how* for defence AI ethics. In moving from principles to a focus on guidelines and tools, this emerging literature on operationalising ethics principles leaves unaddressed crucial intermediate questions about how ethical requirements underpinning the guidance ought to be elicited from AI ethics principles. For example, given their high-level and foundational nature, AI ethics principles are comparable to constitutional principles (Morley et al. 2021), and like constitutional principles, operationalised at a lower level of abstraction the principles are likely to generate tensions requiring balancing in a way not resolvable through recourse to the principles alone (Whittlestone et al. 2019). Guidance should specify, therefore, how AI ethics principles ought to be prioritised in light of the concrete case.

To be effective AI ethics principles need to be coupled with appropriate methodologies to offer domain-specific guidance as to how to apply them (Taddeo and Floridi 2018). See (Taddeo et al. 2024) for such a methodology. It is also important that the normative tradeoffs entailed in the specification of guidelines and the choice of tools are acknowledged. This is because the specification of guidelines in itself entails normative consequences. For example, the metrics chosen for measuring compliance with ethics principles will influence the character of that compliance. The decision to use only quantitative metrics for measuring compliance is likely to generate superficial compliance with the principles (see p.18). Guidance and tools are not, therefore, value-neutral, and the process of designing and specifying guidelines, and of choosing tools, ought to reflect the normativity of the outcomes they generate.

It is not within the scope of this article to recommend specifications for these requirements. These choices ought to be made by the given organisation, through engagement with stakeholders, and in light of the intent and spirit of the AI ethics principles. However, below we outline the key considerations that ought to inform the specification of guidelines and the normative tradeoffs these considerations generate. These features—lifecycle, stakeholders, accountability, auditing, transparency and traceability—are drawn from a qualitative systematic review[4] of domain-agnostic AI governance literature.

Two clarifications are required. First, this is not an exhaustive list of considerations. For instance, context specificity—whether guidelines ought to be specified in a way that applies to particular use cases or broad groupings of use cases—will also be a factor determining outcomes. The context specificity of guidance involves a tradeoff between maximising granularity, to provide detailed ethical instruction on a specific case, and maximising practicality, whereby higher-level guidance provides less detailed instruction for a specific case but greater scope to support standardised and repeatable ethical decision-making across a range of cases. Whilst context specificity is an important determinant of the normative outcomes of guidance, this tradeoff is common to ethical decision-making across all domains and is not uniquely salient to the AI context. This paper focuses on five sets of normative considerations which have particular relevance for the normativity of outcomes in the specification of guidance for AI technologies. Recourse to these considerations can thereby help stakeholders to identify the source(s) of specific problems associated with the application of AI principles into practice. For instance, the inclusion of qualitative as well as quantitative metrics in the audit of AI system performance can help identify whether tensions between principles result from their misapplication or from deeper moral dilemmas requiring deliberation.

Second, whilst defence is a 'high-risk' domain not all applications will be high risk. AI logistics systems for ordering uniforms, say, will not present the same risks as AI in weapon systems, command and control communications, or nuclear retaliation protocols. The considerations outlined below apply to high-level guidance meant for all defence uses cases. However, when eliciting requirements from guidance for particular use cases, the stringency and character of these considerations changes according to the risk magnitude—the combination of ethical harm likelihood and severity—of the use case. For instance, accountability applies both to the use of AI in recruitment processes and in weapon systems, but the way in which accountability is operationalised will differ between the two. Accountability for IHL violations arising from AI in weapon systems will entail international fora in a way not relevant to potential bias arising from AI in recruitment processes. It is beyond

[4] See (Grant and Booth 2009) for an outline of review types.

the scope of this article to develop a risk-based approach for AI principles application. For a methodology to elicit requirements from guidelines according to use-case risk magnitude see (Taddeo et al. 2024).

3 Key methodological considerations for specifying guidelines and assessing their normative tradeoffs

3.1 Lifecycle

Consensus in the relevant literature is that AI ethics guidelines should span the entire lifecycle of an AI system (Alshammari and Simpson 2017; d'Aquin et al. 2018; Leslie 2019; Department of Defense 2022; Cihon et al. 2021; Dunnmon et al. 2021; High-Level Expert Group on Artificial Intelligence 2019; Taddeo et al. 2021; Ayling and Chapman 2022; Mäntymäki et al. 2022). This emphasis on a lifecycle approach tallies with a broader consensus that AI ethics governance must be both holistic and systemic to be effective (Eitel-Porter 2021).

A lifecycle approach mandates the iterative (re)application of principles at successive stages of the project. This is important for reasons of process, product, and purpose (Stilgoe et al. 2013). Regarding process, the needs of a particular project are likely to evolve beyond those originally envisaged at the beginning, and with them new ethical risks may emerge. Regarding product, some AI models, like generative models, can produce new and unexpected behaviours (Taddeo et al. 2022), and therefore ensuring that the product continues to respect ethics principles beyond the release of the product is essential. Regarding purpose, the social and political motivations of a project and the goals or trajectories of innovation may change over time, so ensuring control over the project requires continuous monitoring of its ethical implications.

Evidently, the way the lifecycle is modelled is crucial for specifying effective guidelines. But it also entails choices that are, normatively speaking, far from trivial. This is because the lifecycle of AI, like any technological lifecycle, is a sociotechnical process whereby a "neat theoretical distinction between different stages of technological innovation does not always exist in practice" (La Fors et al. 2019, 210). This makes it problematic to identify the points at which ethical questions are to be asked, certain steps taken, and specific goals met. It also means that stages specified as part of the lifecycle model require justification if they are not to be an arbitrary schema.

The importance of such justifications is illustrated by the choice of lifecycle granularity. Descriptions of the AI lifecycle offered in the relevant literature range from very high-level definitions—referring for example to the three stages of design, develop and deploy—to meticulous descriptions of tasks that each stage of the lifecycle may entail, with each stage further entailing its own sub-stages (Leslie 2019). We argue that care ought to be taken to strike a balance of lifecycle granularity. Too high and too low granularity in the description of the lifecycle is problematic. If too few stages are identified, then the application of principles will not be sufficiently differentiated, leading to blind spots and the creation of ethical risks. Yet, if too many tasks are identified, the iterative application of the principles multiplies, making the guidelines needlessly unwieldy for the user. Too granular a description of the lifecycle also risks being of little value, as it may be outdated quickly by rapid developments in AI that alter the lifecycle stages.[5]

These challenges for modelling the AI lifecycle exist across all domains. However, in the defence domain there are additional challenges that need to be addressed. The first has to do with the interoperability of AI systems. When one nation lends an AI capability to another nation, assurances are likely to be required as to the ethicality (and legality) of that capability given the varying ethical cultures and legal frameworks under which it was designed and developed (Stanley-Lockman 2021). AI lifecycles that are commensurable will be an important aspect of those assurances for interoperability. However, currently there is a fragmentation across the literature on the description of the AI lifecycle across partner defence organisations. This will create inconsistencies and undermine the possibility of creating shared ethical guidelines informing the use of AI. If ethical guidelines for the use of AI in defence are set to become a key requirement for the adoption of this technology and its interoperability amongst allies, it is crucial that the guidelines (and ethical assessment) defined and implemented by partner defence institutions are consistent.

Second, the specification of the AI lifecycle will be important as defence widens the range of vendors for procuring digital technologies (Kinder 2023). This will present particular challenges as AI development will initially be at odds with military capability development and procurement, which tends to be linear and slower-paced, rather than fast-moving and iterative. The specification of a lifecycle is important in that regard for aligning AI development processes and military capability processes with one another.

Some help in finding the right granularity to model the AI lifecycle may be provided by technical standards and frameworks. In this case, however, issues concerning scope and purpose emerge. A system may be modelled in different ways (in terms of granularity and scope) depending on the

[5] For example, consider generative AI models may be used to write code for AI, or how the use of synthetic data may impact steps of the lifecycle, like data collection and labelling.

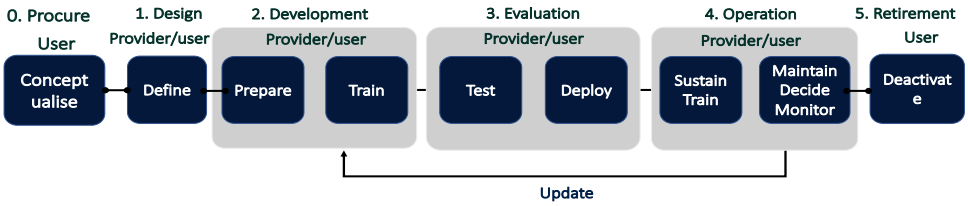


Fig. 3 The AI lifecycle modelled using the highly-granular GoAEthics. Adapted from ISO/IEC/IEEE 12207:2017. See: (Taddeo et al. 2024. This model is adapted from (Floridi et al. 2022))

purpose of the model. Standards and frameworks not developed with the aim of aiding ethical analysis or the application of ethics principles, may prove to be inadequate if not misleading. We propose the following lifecycle model as illustrating the desirable level of granularity for understanding where decisions can lead to unethical consequences (see Fig. 3).[6]

3.2 Stakeholders

There is increasing recognition in the literature that the ethical governance of AI is a multi-stakeholder phenomenon, and that stakeholder involvement is necessary for the specification of AI ethics guidelines (Alshammari and Simpson 2017; Gasser and Almeida 2017; Butcher and Beridze 2019; Jobin et al. 2019; Leslie 2019; Krafft et al. 2020; Hickok 2021; Metcalf et al. 2021; Seppälä et al. 2021; Stix 2021; Ayling and Chapman 2022; Georgieva et al. 2022). Stakeholder involvement is important for effective interpretation of principles as a diversity of perspectives can be instrumental to refining the elicitation of requirements from the principles. Stakeholder involvement also has profound ethical implications, seen as one aspect for maintaining democratic norms and values as AI technologies become more embedded across society. Indeed, the call for multi-stakeholder participation is part of a growing recognition that AI ethics should be concerned with issues of political and social justice (Fukuda-Parr and Gibbons 2021). Some authors argue that involving a wider range of stakeholders in the operationalisation of ethics principles is a corrective to the fact that most sets of AI principles were developed in the global north (Hickok 2021). This is particularly important given that ethics principles can be interpreted in different ways across different groups and contexts (Whittlestone et al. 2019).

The notion of stakeholder 'involvement' has two senses underpinning different approaches to the inclusion of stakeholders in the normative processes, each with significantly different normative outcomes. On the one hand, involving stakeholders can mean giving adequate recognition to the range of stakeholders affected by AI and their interests. In this regard, the specification of AI ethics guidelines can build upon the existing stakeholder theory, which has a well-established framework for identifying stakeholders, describing their interests, and assigning, if any, their responsibilities (Donaldson and Preston 1995; Ayling and Chapman 2022). This type of involvement may also take a consultative approach, whereby ethical analysis is undertaken after an explicit engagement with stakeholders e.g., via a workshop, focus group, or deliberative mini-public. With consultative approaches, stakeholders are not directly involved in ethical analysis, and resulting guidelines are justified on the basis of the coherence of the proposed solutions with the adopted moral theory (Davies et al. 2015). Some consider the need to identify stakeholders in this way as important for stakeholder communication and generating trust (Seppälä et al. 2021), for identifying and mitigating the harms and risks introduced by AI systems, and for providing recompense and explanation to affected parties where harms occur (Georgieva et al. 2022).

On the other hand, 'involvement' can also mean the active participation of stakeholders across the AI lifecycle in the specification of ethical guidelines. This requires that stakeholders participate in the analysis and reaching of normative conclusions (Davies et al. 2015). This characterises the dialogical approach to ethical analysis, which is premised on the idea that consensus-based methods justify normative conclusions (Widdershoven et al. 2009). Some dialogical approaches rely on the idea that dialogue can lead to individuals reaching a shared understanding of the world, leading to agreement on the correct solution. Other interpretations argue that democratic authority rather than shared interpretation and consensus provides normative justification (Kim et al. 2009). In this case, justification flows from the legitimacy of the democratic process invoked to draw the conclusions, rather than from the actual outcome or solution. Here, dialogue is seen not as,

[6] We note that some approaches to modelling the AI lifecycle differentiate lifecycle stages via role-defined actors. This is the approach taken by the European Commission's High-Level Expert Group model of the AI lifecycle (High-Level Expert Group on Artificial Intelligence 2019, 14). This creates three problems: the first is that it does not tell the actor whether (and if so how) the stage under their remit its itself differentiated. The second is that it runs contrary to a recommendation found often in the literature that P2P at each stage of the lifecycle requires multidisciplinary teams (Leslie 2019, 5). Third, there may be stages in the lifecycle, such as planning (e.g. deciding on justified use, deciding success metrics) that do not have a corresponding role-defined actor.

> […] an instrument or technique to reach better decisions; it is rather understood as an ongoing, social learning process in which participants develop new and richer understandings of their practice (Widdershoven et al. 2009, 239)

Both the consultative and dialogical approaches offer important insights when considering the specification of ethical guidelines in the defence domain. For example, a dialogical approach for identifying ethical risks and solutions is key when developing ethical guidelines for AI systems which will impact various stakeholders differently. However, there are feasibility tradeoffs which can mitigate against the viability of the dialogical approach. The dialogical approach has been brought into debate on the ethical governance of AI through bioethics research methodologies. The appeal of bioethics methodologies for the ethical governance of AI stems from the fact that bioethics has been addressing ethical challenges related to the use of new technologies and their impact on fundamental rights for decades (Beauchamp and Childress 2013). However, the broad societal scope of AI technologies and their impacts, their variable purposes, institutional, accountability and application contexts can pose challenges to the consistent application of methodologies which have developed or matured in the narrower, specific context of bioethics.

Deciding on the feasible scope of a dialogical approach is important since the range of stakeholder involvement is relative to the politico-normative goals of that involvement. Where the goal is to ensure reliable and safe systems, and ones that are compliant with the existing regulation, the range of stakeholders is likely to be narrower (Sanderson et al. 2022, 2). Where the aims of stakeholder participation have to do with public interest and issues of justice, the range of stakeholders participating is likely to widen (Winfield and Jirotka 2018, 4; High-Level Expert Group on Artificial Intelligence 2019, 19). However, as the range of stakeholders involved in the specification of guidelines grows, so too do the practical difficulties of involving them in a meaningful way, thereby potentially limiting the capacity to ensure effective guidelines. Additional feasibility problems arises from the sensitive nature of digital technology design and development in the defence domain. The dialogical approach inherits from hermeneutics the assumption that a fuller understanding of a morally complex situation requires an articulation and exploration of the various (sometimes conflicting) stakeholder perspectives about that situation. This in turn presupposes a threshold of information about the situation in order to "discern what matters" (Widdershoven et al. 2009, 239).

We recommend that stakeholder feasibility be assessed along four fronts:

- *How are stakeholders classified?* e.g., technical vs non-technical stakeholders (Stix 2021).
- *How are stakeholders to participate?* What mechanisms enable stakeholder participation and elicit feedback?
- *When are stakeholders to participate?* Would, for instance, consultations be required every time an AI system undergoes a substantial upgrade or put to different uses/different contexts?
- *Are there tradeoffs?* There are likely to be tradeoffs whereby public interest requires reducing stakeholder participation, such as in the national defence and security domains.
- *Are they accountable?* To what extent do stakeholders participating in the operationalisation of principles remain accountable for ethical risks, shortcomings, and mistakes that occur after the implementation of the ethical guidelines?

These questions remain largely unaddressed in the relevant literature, and if they remain so there is a risk that multi-stakeholder approaches to the specification of ethical guidelines may lead to superficial involvement by stakeholders or cumbersome processes that may hinder the development of effective and agile guidelines.

3.3 Accountability

Given the formative nature of AI governance, it will be important in guidance for organisations to define (at minimum) the process of accountability. Accountability mechanisms that span the whole of the AI lifecycle are important for enabling human oversight of the system, for identifying and holding human actors to account when obligations are not met, and for disincentivising non-compliance with ethical principles. However, realising accountability for AI systems has its own challenges. The opaque and unpredictable outcomes of AI systems makes it difficult to identify those accountable (Tsamados et al. 2021; Taddeo et al. 2022). This can lead to accountability gaps, as well as to false perceptions of relevant stakeholders with respect to their answerability for the outcomes of an AI system (Novelli et al. 2023). Therefore, the failure to establish lines of accountability can lead to 'ethics shirking', with harms generated by AI systems pushed downstream.

Two clarifications are in order here. The first can be dealt with swiftly: whilst often conflated 'responsibility' and 'accountability' are distinct concepts. Accountability is scrutiny from an external point of view and is a form of 'answerability', whilst (moral) responsibility is an internal point of view, i.e. an assessment of agency (Novelli et al. 2023, 3). This is significant because closing the AI accountability gap, for instance through institutional processes of scrutiny, does not mean that the moral

responsibility gap generated by AI is closed as well (Taddeo and Blanchard 2022).

The second clarification refers to the understanding of accountability, which impacts the way in which ethics principles mandating accountability are operationalised. Accountability is a relation of answerability involving "an obligation to inform about and justify one's conduct to an authority" (Novelli et al. 2023, 3). This relation presupposes three conditions: authority recognition, interrogation, and limitation of power. The content of this relation is determined by different practices, values, and measures—the 'accountability regime'. Novelli et al. identify seven features of an accountability regime: *context* (what for?), *range* (about what?), *agent* (who is accountable?), *forum* (to whom is an account due?), *standard* (according to what?), *process* (how?), and *implications* (what follows?).

How these features are determined will represent a compromise between the sociotechnical system entailing the AI and the goals of accountability. To this end, both the sociotechnical system and the goals of accountability need to be made explicit. Regarding sociotechnical systems, they presuppose their own expectations, roles, procedures, cultural backgrounds, and coordination mechanisms that inform the accountability regime (Theodorou and Dignum 2020). For example, defence organisations will have their own accountability structures, and, with parsimony being desirable, should rely on existing good governance when developing AI accountability practices.

As for the goals, Novelli et al. (2023, 6) identify four types of accountability goals: *compliance* aims to "bind the agent to align with ethical and legal standards." *Report* aims to "ensure that the agent's conduct is properly recorded to explain and justify it to the forum," thus enabling the forum to "challenge and disapprove the agent's conduct." *Oversight* aims to examine information, obtain evidence, and evaluate the agent's conduct according to the rules of the deployment context. *Enforcement* aims to "determine what consequences the agent must bear […] according to the evidence gathered during the report and oversight."

Depending on the goal, operationalising accountability will require different tools. For example, when considering ethical principles, accountability ought to serve the goal of compliance with those principles. Thus it requires measures, practices, and tools suited to binding agents to a given set of ethical principles. Some argue that auditing is the best tool to foster accountability for compliance (Sandvig et al. 2014; Raji et al. 2020; Mökander and Floridi 2021b; Mökander et al. 2021; Costanza-Chock et al. 2022). Others propose self-assessments; ethics advisory panels with power to veto projects that do not adhere to ethics guidelines (Kroll 2018; Theodorou and Dignum 2020, 11; Seppälä et al. 2021, 3; Morley et al. 2021, 252); impact assessments (Schiff et al. 2020); and participatory design methods (Mäntymäki et al. 2022, 604).

We submit that, since the same tool can serve different accountability goals, the issue is not one of choosing between different tools per se, but which accountability goal is best to foster ethical AI in a specific domain and therefore should be operationalised. For example, some recommend impact assessments for closing the accountability gap (Reisman et al. 2018; Schiff et al. 2020; Ada Lovelace Institute 2022), but impact assessments can be used both *ex-ante* or *ex-post*, depending on their design and the point of use in the lifecycle.

The choice between different goals and tools for accountability is not value-neutral. *Ex-post* algorithmic assessments that disclose (report) information about the use of AI systems and their impacts primarily serve values such as transparency and justification, and cultivating public trust, rather than focussing on achieving the goal of compliance. Whereas, *ex-ante* auditing processes can cultivate value alignment but put less weight on values such as transparency and explainability. Thus when operationalising the ethical principle of accountability, a decision as to whether this should be implemented *ex-ante* or *ex-post* (or both) should be made. The decision may differ depending on how other requirements for specifying the guidelines are chosen.

3.4 Auditing

There is growing consensus on the need for auditing to achieve ethical governance of AI (henceforth 'ethics-based auditing', EBA)(Mökander and Floridi 2021a). In the defence domain, the UK Ministry of Defence (2022, 10) has highlighted audits as important for compliance with the principle of 'understanding', whilst the US DoD identifies auditing AI systems as essential for realising the principle of traceability (DIB 2019). The Defence Innovation Unit (DIU) also foregrounds the importance of AI auditing in its guidelines, particularly the need to establish clear plans for routine system auditing as well as roles and responsibilities for those audits (Dunnmon et al. 2021, 9, 15). NATO has also stated that auditing will be required for putting its ethics principles into practice (Stanley-Lockman and Christie 2021).

There are three themes characterising the literature on EBA: scope, including which part of the lifecycle and which stakeholders to involve; metrics; and procedures. The debate on the scope of EBA hinges on the question as to whether auditing processes should be proactive and iterative. There is no consensus around this topic. For example the US DIU recommends establishing "plans for routine system auditing" at the development phase of the lifecycle, but it then considers running auditing processes only at the deployment stage (Dunnmon et al. 2021, 4). In its 'Responsible Artificial Intelligence Strategy and Implementation Pathway',

the DoD (2022, 14) requires "data audit at design assessment stage." Several documents state a commitment to a range of stakeholder interests. A holistic approach to EBA requires identifying a wide range of stakeholders to assess whether effects of AI systems will be commensurate with specified principles. Thus, questions posed in Sect. 3.2 are also relevant here.

The choice of the metrics used for the assessment has an impact on the outcome of the audit (Costanza-Chock et al. 2022). In defence guidelines there is little discussion of the methods used for auditing processes, but where preference is shown it is for quantitative metrics.[7] As the DIU guidelines state:

> Models can be audited in multiple ways, ranging from internal code and training process reviews to fuzzing and deterministic testing, and different applications will require different degrees of capability auditing (Dunnmon et al. 2021, 29).

Quantitative metrics that assess performance and system robustness are important for anticipating system outcomes, itself important for human control over the system. However, they also have important limitations. They abstract from the AI sociotechnical system, which is problematic because principles do not have meaning outside of context, and quantitative metrics cannot track the way principles like 'fairness' vary between contexts and stakeholders. Quantitative metrics may also obscure the tensions that arise between principles and may create an undue focus on meeting the measures themselves rather than helping or aiding human judgement to envisage real-world risks and harms arising from the use of an AI system. EBA requires drawing:

> […] the boundaries of abstraction to include people and social systems as well, such as local incentives and reward structures, institutional environments, decision-making cultures and regulatory systems (Selbst et al. 2019, 64).

This requires balancing the range of tools used to avoid privileging quantitative metrics.

What an organisation decides with respect to the scope, the metrics, and the procedures of EBA, will have important implications for the outcome of the audit itself. Whilst focus and consensus on the need for EBA have grown in the recent years, little attention has been dedicated to developing and testing such methodologies. When considering EBA from a procedural point of view, this can be either internal or external. External audit entails the use of external agents to interrogate the results as impartially as possible. As Raab (2020, 13) writes, "an organisation […] that simply marks its own homework cannot make valid claims to be trustworthy." To mitigate this risk, independent parties may sit on organisational ethics boards tasked with undertaking audits; organisations may publish guidelines and case studies of ethical risk assessments for scrutiny by external experts; and may also produce annual transparency reports. Some literature (Raji et al. 2020; Centre for Data Ethics and Innovation 2021) supports the use of internal auditors insofar as they are likely to have greater access to information about the expected functioning of the system and its effects. This is particularly important for considering a whole lifecycle approach. The choice between internal and external procedures should be informed by the risks related to the use of AI systems as well as the duty to ensure adherence to the ethical principles of an organisation.

3.5 Transparency and traceability

The relevant literature identifies two types of transparency: system transparency and systems development transparency (Vakkuri and Kemell 2019, 4; Ryan and Stahl 2021, 66; Seppälä et al. 2021, 3). System transparency refers to the transparency of the AI product itself, including components such as data or the algorithms. Systems development transparency refers to the transparency of research and innovation processes leading to the AI product. This second type of transparency is also understood as traceability. It requires establishing

> […] not only how a system worked but how it was created and for what purpose, in a way that explains why a system has particular dynamics or behaviours (Kroll 2021, 758).

Both types of transparency are important for understanding an AI system, but system development transparency is key to operationalising AI ethics principles, because it allows for an understanding of the decisions (what, why, and by whom) made during the design and development of AI systems and their conformity with broader governance goals.

When considering system development transparency, two measures are important. The first concerns auditing, as we have focussed on auditing in the previous section, we shall not consider this measure here. The second measure is a commitment to replicability (Morley et al. 2020, 2160). As Sanderson et al. (2022, 6) write, the operationalization of AI ethics principles will require.

> […] the application of reusable design methods, where the design and implementation of system components,

[7] Quantitative metrics include: checking training data appropriateness for modelling, assessing data representativeness, assessing bias in input data, and measuring accuracy of the AI system on individual subgroups (Costanza-Chock et al. 2022).

as well as their integration, follows known patterns […].

Transparency also concerns the process used for the specification of ethical guidelines, in order to foster accountability and trust. For example, the operationalisation of system transparency and transparency of system development requires striking a balance between algorithmic transparency, information sensitivity requirements, and military necessity. It may be that specific information cannot be disclosed, but this does not mean that the need for transparency is redundant. For example, disclosing the criteria with which this balance is achieved would offer some assurance as to how ethical risks are addressed, fostering accountability of, and trust in, the institution making this decision.

It is important to stress the need for a model of "good transparency" (Porter et al. 2022), one that fosters access to relevant information to identify ethical risks, points of intervention, and accountability. To this end, Porter et al. (2022) cite Grice's (1975) four maxims of cooperative communicative exchange: quantity (the right amount of information is conveyed), quality (it is truthful), relevance (the information is salient), and manner (its transmission facilitates effective exchange of information and understanding). In the defence domain, achieving good transparency may require balancing these four elements and may vary with the recipient of the information.

4 A pro-ethical institutional culture

The above describes key considerations to be navigated when specifying guidance for the implementation of principles into practice. Since each can generate varying ethical outcomes, each entails tradeoffs that will have to be navigated by the organisation. Deciding on these tradeoffs will require recourse to the spirit of the ethics principles as they were intended when designed. However, all of these points for consideration are moot if the organisational culture does not itself enable serious reflection and judgement about these considerations, and does not support the ethical governance of AI. To achieve this result, and to avoid risks of malpractice, these efforts need to take place within a pro-ethical institutional culture, where ethics is not perceived as an add-on, or treated as an extra burden for practitioners, but is a constitutive, non-negotiable element of everyday practice, which enables the achievement of positive results. If the pressing choices raised by AI governance are dealt with absent the wider organisational (and societal) context, or in contexts without strong pre-existing cultures of organisational responsibility, they could be at risk of being gamed, facilitating malpractices, or reducing AI ethics to a tick-box exercise to justify the existing decisions, rather than supporting responsible and ethical decision-making. In this sense, the existing issues around organisational governance, employee rights, labour policies, as well as means to challenge AI-based decisions, must also be in place to underpin navigating the above tradeoffs. This will also need to be coupled with ethics training to foster amongst practitioners awareness of ethical risks, problems, complexities, and opportunities associated with the use of AI. Indeed, vigilance over organisational culture and practices capable of identifying whether AI systems are developed and used in ways consistent with norms will be required for the fact that the diffusion of AI systems will have a disruptive effect on such norms and labour practices.

This applies as much to defence organisations as to defence vendors. We noted above the likelihood of friction between defence and digital technology vendors given the inconsistencies between military capability development and the AI development lifecycle. Crucial here will be third-party ethics-based audits to ensure that vendors undertake AI projects in a way that is consistent with Defence AI ethics principles, as well as broader norms and values. Indeed, the DoD DIU warns that it is "a red flag if the vendor refuses to allow third-party or government system audits without a very compelling reason" (Dunnmon et al. 2021, 29). Here the DIU refers to the type of audits for verifying system outputs, but serious appetite for ethical compliance ought also to take refusals for ethics-based auditing as a red flag. Further research should be undertaken on realising third-party ethics-based audits in defence whereby the sensitive nature of defence AI projects raises audit transparency challenges.

5 Conclusion

In this article we have outlined the key features for informing the specification of AI ethics guidelines and the normative tradeoffs encompassed by these features. There are five areas where important normative tradeoffs are made: (1) the model of the AI lifecycle; (2) the scope of stakeholder involvement; (3) the accountability goals chosen; (4) the choice of auditing requirements; (5) the choice of mechanisms for transparency and traceability mechanisms. These were identified through a systematic review of AI governance literature. For each of the five areas, we have provided initial recommendations for navigating the tradeoffs entailed by choices made in the specification of guidance for implementing AI ethics principles.

Acknowledgements The authors are grateful for feedback given by participants at the May 2023 'Expert workshop for applying AI ethics principles in the defence domain', co-hosted by The Alan Turing Institute and The Oxford Internet Institute. The authors would also like to thank James Rosie for helpful comments on the draft.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Funding Alexander Blanchard, Christopher Thomas and Mariarosaria Taddeo have been funded by the Dstl Ethics Fellowship held at the Alan Turing Institute. The research underpinning this work was funded by the UK Defence Chief Scientific Advisor's Science and Technology Portfolio, through the Dstl Artificial Intelligence Programme, grant number R-DST-TFS/D2013. This paper is an overview of UK Ministry of Defence (MOD) sponsored research and is released for informational purposes only. The contents of this paper should not be interpreted as representing the views of the UK MOD, nor should it be assumed that they reflect any current or future UK MOD policy. The information contained in this paper cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

Data availability Not applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ada Lovelace Institute (2022) algorithmic impact assessment: a case study in healthcare. London: Ada Lovelace Institute. https://www.adalovelaceinstitute.org/wp-content/uploads/2022/02/Algorithmic-impact-assessment-a-case-study-in-healthcare.pdf. Accessed 18 Dec 2023

Alshammari M, Simpson A (2017) Towards a Principled Approach for Engineering Privacy by Design. In: Schweighofer E, Leitold H, Mitrakas A, Rannenberg K (eds) Privacy technologies and policy. Lecture Notes in Computer Science, vol 10518. Springer International Publishing, Cham, pp 161–177. https://doi.org/10.1007/978-3-319-67280-9_9

Aquin M, Troullinou P, O'Connor NE, Cullen A, Faller G, Holden L (2018) Towards an "Ethics by Design" methodology for AI research projects. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp 54–59. New Orleans LA USA: ACM. https://doi.org/10.1145/3278721.3278765.

Ayling J, Chapman A (2022) Putting AI ethics to work: are the tools fit for purpose? AI and Ethics 2(3):405–429. https://doi.org/10.1007/s43681-021-00084-x

Beauchamp TL, Childress JF (2013) Principles of biomedical ethics, 7th edn. Oxford University Press, New York

Blanchard A, Taddeo M (2023) The ethics of artificial intelligence for intelligence analysis: a review of the key challenges with recommendations. Digital Society 2(1):12. https://doi.org/10.1007/s44206-023-00036-4

Butcher J, Beridze I (2019) What is the state of artificial intelligence governance globally? The RUSI Journal 164(5–6):88–96. https://doi.org/10.1080/03071847.2019.1694260

Canca C (2023) AI ethics and governance in defence innovation: implementating AI ethics framework. In: Raska M, Bitzinger RA (eds) The AI wave in defence innovation: assessing military artificial intelligence strategies, capabilities, and trajectories. Routledge, London

Centre for Data Ethics and Innovation (2021) The role of independence in assuring AI. AI Assurance Guide. 2021. https://cdeiuk.github.io/ai-assurance-guide/independence/

Cihon P, Schuett J, Baum SD (2021) Corporate governance of artificial intelligence in the public interest. Information 12(7):275. https://doi.org/10.3390/info12070275

Coldicutt R, Miller C (2019) People, power, and technology: the tech workers' View'. London: Doteveryone. https://doteveryone.org.uk/wp-content/uploads/2019/04/PeoplePowerTech_Doteveryone_May2019.pdf. Accessed 18 Dec 2023

Costanza-Chock S, Raji ID, Buolamwini J (2022) Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem'. In: 2022 ACM Conference on fairness, accountability, and transparency, pp 1571–83. Seoul Republic of Korea: ACM. https://doi.org/10.1145/3531146.3533213

Davies R, Ives J, Dunn M (2015) A systematic review of empirical bioethics methodologies. BMC Med Ethics 16(1):15. https://doi.org/10.1186/s12910-015-0010-3

Department of Defense (2022) Responsible artificial intelligence strategy and implementation pathway. Department of Defense

Devitt K, Michael G, Scholz J, Bolia R (2020) 'A Method for Ethical AI in Defence.' DSTG-TR-3786. Australian Department of Defence, Canberra

DIB (2019) AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. Defense Innovation Board

DIB (2020) AI principles: recommendations on the ethical use of artificial intelligence by the Department of Defense-Supporting Document. Defense Innovation Board [DIB]. https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF

DoD Responsible AI Working Council (2022) Responsible artificial intelligence strategy and implementation pathway

Donaldson T, Preston LE (1995) The Stakeholder Theory of the Corporation: Concepts, Evidence, and Implications. Acad Manag Rev 20(1):65–91

Dunnmon J, Goodman B, Kirechu P, Smith C, Van Deusen A (2021) Responsible AI guidelines in practice: operationalizing DoD's ethical principles for AI. California: Defense Innovation Unit. https://assets.ctfassets.net/3nanhbfkr0pc/acoo1Fj5uungnGNPJ3QWy/3a1dafd64f22efcf8f27380aafae9789/2021_RAI_Report-v3.pdf. Accessed 18 Dec 2023

Eitel-Porter R (2021) Beyond the promise: implementing ethical AI. AI Ethics 1(1):73–80. https://doi.org/10.1007/s43681-020-00011-6

Floridi L (2019) Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol 32(2):185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harv Data Sci Rev. https://doi.org/10.1162/99608f92.8cd550d1

Floridi L, Holweg M, Taddeo M, Silva JA, Mökander J, Wen Y (2022) capAI - a procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. SSRN Electron J. https://doi.org/10.2139/ssrn.4064091

Fors La, Karolina BC, Keymolen E (2019) Reassessing values for emerging big data technologies: integrating design-based and application-based approaches. Ethics Inf Technol 21(3):209–226. https://doi.org/10.1007/s10676-019-09503-4

Fukuda-Parr S, Gibbons E (2021) Emerging consensus on "Ethical AI": human rights critique of stakeholder guidelines. Global Pol 12(S6):32–44. https://doi.org/10.1111/1758-5899.12965

Gasser U, Almeida VAF (2017) A layered model for AI governance. IEEE Internet Comput 21(6):58–62. https://doi.org/10.1109/MIC.2017.4180835

Georgieva I, Lazo C, Timan T, Fleur A, van Veenstra. (2022) From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI and Ethics 2(4):697–711. https://doi.org/10.1007/s43681-021-00127-3

Grant MJ, Booth A (2009) A typology of reviews: an analysis of 14 review types and associated methodologies: a typology of reviews, Maria J. Grant Andrew Booth. Health Inform Libr J 26(2):91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Grice HP (1975) Logic and conversation. In: Davidson D and Harman G (eds) The logic of grammar, pp 64–75

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Mind Mach 30(1):99–120. https://doi.org/10.1007/s11023-020-09517-8

Hajer M (2003) Policy without polity? Policy analysis and the institutional void. Policy Sci 36(2):175–195. https://doi.org/10.1023/A:1024834510939

Hickok M (2021) Lessons learned from ai ethics principles for future actions. AI Ethics 1(1):41–47. https://doi.org/10.1007/s43681-020-00008-1

High-Level Expert Group on Artificial Intelligence (2019).'Ethics guidelines for trustworthy AI. Brussels: European Comission. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf. Accessed 18 Dec 2023

Jobin A, Ienca M, Vayena E (2019) The Global Landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399

Kim SYH, Wall IF, Stanczyk A, De Vries R (2009) Assessing the public's views in research ethics controversies: deliberative democracy and bioethics as natural allies. J Empir Res Hum Res Ethics 4(4):3–16. https://doi.org/10.1525/jer.2009.4.4.3

Kinder T (2023) silicon valley chiefs urge pentagon procurement overhaul. Financial Times, 26 June 2023, sec. Tech start-ups. https://www.ft.com/content/45da39f2-4e05-46f1-96f4-813fbba79b16. Accessed 18 Dec 2023

Krafft T, Hauer M, Fetic L, Kaminski A, Puntschuh M, Otto P, Hubig C et al (2020) from principles to practice - an interdisciplinary framework to operationalise AI ethics. AI Ethics Impact Group

Kroll JA (2018) Data science data governance [AI Ethics]. IEEE Secur Priv 16(6):61–70. https://doi.org/10.1109/MSEC.2018.2875329

Kroll JA (2021) Outlining traceability: a principle for operationalizing accountability in computing systems. In: Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency, 758–71. FAccT '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445937.

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, London

Mäntymäki M, Minkkinen M, Birkstedt T, Viljanen M (2022) Defining organizational AI governance. AI and Ethics 2(4):603–609. https://doi.org/10.1007/s43681-022-00143-x

McNamara A, Smith J, Murphy-Hill E (2018) Does ACM's Code of ethics change ethical decision making in software development?' In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the foundations of software engineering, pp 729–33. Lake Buena Vista FL USA: ACM. https://doi.org/10.1145/3236024.3264833.

Metcalf J, Moss E, Watkins EA, Singh R, Elish MC (2021) Algorithmic Impact assessments and accountability: the co-construction of impacts. In: Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency, pp 735–46. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445935.

Ministry of Defence (2022) Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in defence. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082991/20220614-Ambitious_Safe_and_Responsible.pdf. Accessed 18 Dec 2023

Mökander J, Floridi L (2021a) Ethics-based auditing to develop trustworthy AI. Mind Mach 31(2):323–327. https://doi.org/10.1007/s11023-021-09557-8

Mökander J, Morley J, Taddeo M, Floridi L (2021) Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. Sci Eng Ethics 27(4):44. https://doi.org/10.1007/s11948-021-00319-4

Morley J, Floridi L, Kinsey L, Elhalal A (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26(4):2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L (2021) Ethics as a service: a pragmatic operationalisation of AI ethics. Mind Mach 31(2):239–256. https://doi.org/10.1007/s11023-021-09563-w

Munn L (2022) The uselessness of AI ethics. AI & Soc. https://doi.org/10.1007/s43681-022-00209-w

Novelli C, Taddeo M, Floridi L (2023) Accountability in artificial intelligence: what it is and how it works. AI & Soc. https://doi.org/10.1007/s00146-023-01635-y

Peters D (2019) Beyond principles: a process for responsible tech. The Ethics of Digital Experience (blog). 14 May 2019. https://medium.com/ethics-of-digital-experience/beyond-principles-a-process-for-responsible-tech-aefc921f7317. Accessed 18 Dec 2023

Porter Z, Habli I, McDermid J, Kaas M (2022) A principles-based ethical assurance argument for AI and autonomous systems. arXiv. https://doi.org/10.48550/arXiv.2203.15370

Raab CD (2020) Information privacy, impact assessment, and the place of ethics. Comput Law Secur Rev 37(July):105404. https://doi.org/10.1016/j.clsr.2020.105404

Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. arXiv. https://doi.org/10.48550/arXiv.2001.00973.

Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic imapct assessments: a practical framework for public agency accountability. AI Now Institute, New York

Ryan M, Stahl BC (2021) Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc 19(1):61–86. https://doi.org/10.1108/JICES-12-2019-0138

Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J, Newnham G, Hajkowicz S, Robinson C, Hansen D (2022) AI ethics principles in practice: perspectives of designers and developers. arXiv. https://doi.org/10.48550/arXiv.2112.07467

Sandvig C, Hamilton K, Karahalios K, Langbort C (2014) Auditing algorithms: research methods for detecting discrimination on internet platforms. Data Discrim Conver Crit Concerns Prod Inquiry 22(2014):4349–4357

Schiff D, Rakova B, Ayesh A, Fanti A, Lennon M (2020) Principles to practices for responsible AI: closing the gap. arXiv. https://doi.org/10.48550/arXiv.2006.04707.

Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on fairness, accountability, and transparency, pp 59–68

Seppälä A, Birkstedt T, Mäntymäki M (2021) From ethical AI principles to governed AI. In: ICIS 2021 Proceedings 10. https://aisel.aisnet.org/icis2021/ai_business/ai_business/10

Stanley-Lockman Z, Christie EH (2021) An artificial intelligence strategy for NATO. NATO Review. 25 October 2021. https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html. Accessed 18 Dec 2023

Stanley-Lockman Z (2021) Responsible and ethical military AI: allies and allied perspectives. Centre for Security and Emerging Technology. https://cset.georgetown.edu/publication/responsible-and-ethical-military-ai/. Accessed 18 Dec 2023

Stilgoe J, Owen R, Macnaghten P (2013) Developing a framework for responsible innovation. Res Policy 42(9):1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Stix C (2021) Actionable principles for artificial intelligence policy: three pathways. Sci Eng Ethics 27(1):15. https://doi.org/10.1007/s11948-020-00277-3

Taddeo M (2013) Cyber security and individual rights, striking the right balance. Philos Technol 26(4):353–356. https://doi.org/10.1007/s13347-013-0140-9

Taddeo M (2015) The struggle between liberties and authorities in the information age. Sci Eng Ethics 21:1125–1138. https://doi.org/10.1007/s11948-014-9586-0

Taddeo M, Floridi L (2018) How AI can be a force for good. Science 361(6404):751–752. https://doi.org/10.1126/science.aat5991

Taddeo M, McNeish D, Blanchard A, Edgar E (2021) Ethical principles for artificial intelligence in national defence. Philos Technol 34(4):1707–1729. https://doi.org/10.1007/s13347-021-00482-3

Taddeo M, Blanchard A (2022) Accepting moral responsibility for the actions of autonomous weapons systems—a moral gambit. Philos Technol 35(3):78. https://doi.org/10.1007/s13347-022-00571-x

Taddeo M, Ziosi M, Tsamados A, Gilli L, Kurapati S (2022) Artificial intelligence for national security: the predictability problem. Centre for Emerging Technology and Security, London

Taddeo M, Blanchard A, Thomas C (2024) From AI ethics principles to practices: a teleological methodology to apply AI ethics principles in the defence domain. Philos Technol. https://doi.org/10.1007/s13347-024-00710-6

Theodorou A, Dignum V (2020) Towards ethical and socio-legal governance in AI. Nat Mach Intell 2(1):10–12. https://doi.org/10.1038/s42256-019-0136-y

Tsamados A, Aggarwal N, Cowls J, Morley J, Roberts H, Taddeo M, Floridi L (2021) The ethics of algorithms: key problems and solutions. AI & Soc. https://doi.org/10.1007/s00146-021-01154-8

Vakkuri V, Kemell K-K (2019) Implementing AI ethics in practice: an empirical evaluation of the RESOLVEDD STrategy. In: Hyrynsalmi S, Suoranta M, Nguyen-Duc A, Tyrväinen P, Abrahamsson P (eds) Software business. Lecture Notes in Business Information Processing, vol 370. Springer International Publishing, Cham, pp 260–275. https://doi.org/10.1007/978-3-030-33742-1_21

Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp 195–200. Honolulu HI USA: ACM. https://doi.org/10.1145/3306618.3314289

Widdershoven G, Abma T, Molewijk B (2009) Empirical ethics as dialogical practice. Bioethics 23(4):236–248. https://doi.org/10.1111/j.1467-8519.2009.01712.x

Winfield AFT, Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos Trans R Soc A Math Phys Eng Sci. https://doi.org/10.1098/rsta.2018.0085