Philosophy & Technology (2024) 37:42
https://doi.org/10.1007/s13347-024-00710-6

RESEARCH ARTICLE

From AI Ethics Principles to Practices: A Teleological Methodology to Apply AI Ethics Principles in The Defence Domain

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

Received: 26 July 2023 / Accepted: 2 February 2024 / Published online: 13 March 2024

Abstract
This article provides a methodology for the interpretation of AI ethics principles to specify ethical criteria for the development and deployment of AI systems in high-risk domains. The methodology consists of a three-step process deployed by an independent, multi-stakeholder ethics board to: (1) identify the appropriate level of abstraction for modelling the AI lifecycle; (2) interpret prescribed principles to extract specific requirements to be met at each step of the AI lifecycle; and (3) define the criteria to inform purpose- and context-specific balancing of the principles. The methodology presented in this article is designed to be agile, adaptable, and replicable, and when used as part of a pro-ethical institutional culture, will help to foster the ethical design, development, and deployment of AI systems. The application of the methodology is illustrated through reference to the UK Ministry of Defence AI ethics principles.

Keywords Artificial Intelligence · Defence Domain · Ethics · Methodology · Military · Practices · Principles

1 Introduction

In this article we set out a methodology to interpret and apply AI ethics principles into practices with a specific focus on the defence domain. In recent years, initiatives and efforts to define domain-dependant AI ethics principles have multiplied (Jobin

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

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

[2] Alan Turing Institute, London, UK

et al., 2019). The defence domain is no exception, the UK (Ministry of Defence, 2022), US (DIB, 2020), and Australian (Devitt et al., 2020) national defence institutions, as well as NATO,[1] have all issued AI ethics principles. However, these principles, like those proposed in other domains, have been criticised as too abstract to offer concrete guidance to the actual design, development and use of AI systems (Coldicutt & Miller, 2019; Munn, 2022; Peters, 2019). As a result, their efficacy to inform decision making has been called into question. For example, a survey including software engineering students and professional software developers showed that.

> "[…] no statistically significant difference in the responses for any vignette were found across individuals who did and did not see the code of ethics, either for students or for professionals" (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 national defence and security (Taddeo, 2013, 2014; Blanchard & Taddeo, 2023).

, administration of justice (Angwin et al., 2016), and healthcare (Obermeyer et al., 2019), where ethical risks related to the use of AI systems are severe and may put individual rights and democratic values under sharp devaluative pressure. This is also because their generic nature may lead to AI ethics principles being.

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

This, in turn, may reduce AI ethics efforts to façade operations, voided of any concrete outcomes and induce malpractices (Floridi, 2019, 186).

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 & Floridi, 2018), to answer the question 'what do I (provider/designer/developer/user) have to do to ensure that this step of the AI lifecycle respects the AI ethics principles?' This has motivated a shift from the *what* to the *how* in AI ethics research (Floridi, 2019, 185), referred to as the "third wave of scholarship on ethical AI" (Georgieva et al., 2022). As a result, there is a growing body of literature focused on developing AI ethics tools and processes to implement AI ethics principles.[2]

However, by focusing directly on tools and applicable solutions, the third wave literature leaves unaddressed crucial, normative questions concerning the interpretation of these principles. The choice of AI ethics tools often assumes normative decisions, which cannot be made by referring to the principles alone (Blanchard et al., 2024). For example, the choice of ethics-based auditing tools implies a decision as to the type of metrics, whether qualitative or quantitative, to be used for the assessment. In turn, the type of metric determines the scope of the audit and of the ethical assessment. This poses the need for guidance as to what type of metrics one should choose in specific situations. Given their high-level

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

[2] Morley et al., (2020a, 2020b) have compiled a taxonomy of these tools.

and foundational nature, these principles have been compared to constitutional principles (Morley et al., 2021). Like constitutional principles, AI ethics principles can be interpreted following different methodologies, but it is still unclear which methodology works best for such principles and whether there are domain-dependent aspects requiring the use of different methodologies in different domains. At the same time, when applied to specific cases, AI ethics principles may generate tensions requiring balancing of the principles, the correct balance cannot be identified through recourse to the principles or tools alone (Whittlestone et al., 2019). The question emerges as to what kind of criteria should shape such a balancing.

In short, there is a midway step between AI ethics principles and AI ethics tools which concerns the specification of a methodology for the interpretation and application of these principles into practices. This step has been disregarded in the relevant literature. Closing this gap is the goal of this article. To do so, we describe a new, three-step methodology for the interpretation and application of AI ethics principles in the defence domain.[3] The methodology is designed to be used by a multi-stakeholder ethics board which follows the steps to extract ethical guidelines from high-level principles. The three steps are abstraction, ethical requirements elicitation, and balancing.

To illustrate the application of the methodology we use the ethics principles for AI provided by the UK MoD (Ministry of Defence, 2022).[4] However, it is worth stressing that the proposed methodology is principle-agnostic, and may be applied to any set of AI ethics principles. It is designed to address the use of AI in high-risk domains, like defence and security, healthcare, and the administration of justice, by ensuring representation and involvement of all relevant stakeholders (including impacted stakeholders) and mitigating measures for risks of conflict of interests and ethics devolution.

In the rest of this article, we introduce the ethics board and the three-step methodology in Sect. 2, offer an example of the application of the methodology in Sect. 3, and conclude our analysis in Sect. 4.

2 A Three-step Methodology to Extract Guidelines from Ethical AI Principles in Defence

Three categories of risks are particularly relevant when specifying ethical guidelines for public institutions working in high-risks domains. The first risk concerns the moral legitimacy of the resulting guidelines – i.e. whether the guidelines mirror correctly the AI ethics principles of an institution as well as the overarching democratic

[3] For a review of relevant literature that provides the justification for the methodology see Blanchard et al., (2024).

[4] These principles were announced in the 'Ambitious, Safe, and Responsible' document as part of the 2022 Defence AI Strategy.

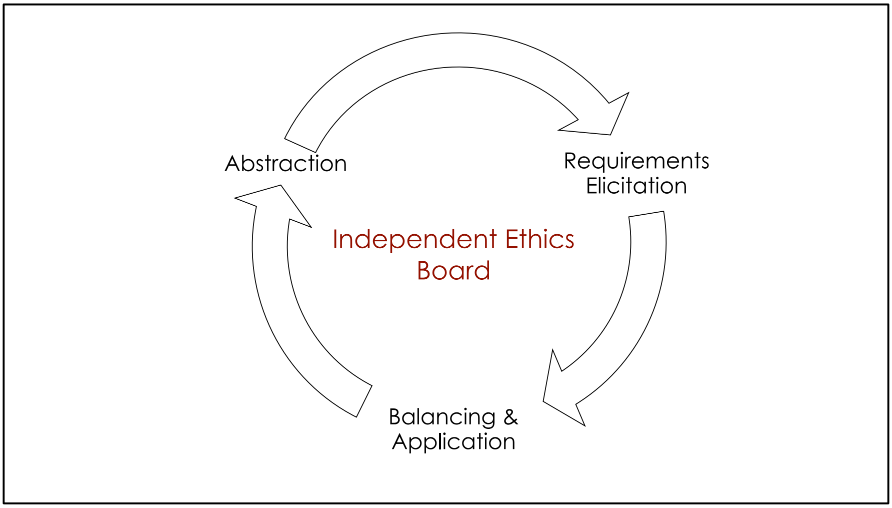


Fig. 1 The three-step methodology for the definition of ethical guidelines for AI in defence

values and respect for human rights; second, the risk of ethics devolution,—i.e. where the burdens for implementing AI ethics principles are shifted from institutions to employees; third, risks to limited reproducibility and scrutability of the process used to interpret principles to extract guidelines. The methodology, which centres on an independent, multistakeholder ethics board (EB), is set to address all three categories of risks.

The EB follows a three-step process to extract ethical guidelines from high-level principles. The three steps are (1) identifying an appropriate level of abstraction (LoA) for modelling the AI lifecycle; (2) interpreting the principles to extract specific requirements to be met at each stage of the AI lifecycle; and (3) defining the criteria to inform purpose and context-specific balancing of the principles. The methodology is an iterative process that is refined while being implemented. It is important to stress that this process does not happen in a vacuum, but is influenced by, and strives to be consistent with, the rights and values already underpinning the work of defence institutions in democratic societies (Fig. 1). We outline the EB before describing each of the three steps.

2.1 Independent, Multistakeholder Ethics Board

The EB has three tasks: identifying the LoA to model the AI lifecycle; interpreting the principles to extract specific requirements to be met at each step of the AI lifecycle; and defining criteria to inform purpose- and context-specific balancing of the principles. These tasks require in-depth expertise on AI, AI ethics, military ethics, Just War Theory, national defence and security, as well as expertise on civil rights, democratic values, relevant international law (including International Humanitarian Law and International Human Rights Law), and the internal procedures of the relevant defence institution. Thus, the EB should include experts on all these areas.

Aside from the breadth of expertise, the EB must be independent from the institution adopting the principles and include representatives of the different categories of stakeholders impacted by the use of AI in defence.

Three reasons support the need for the independence of the EB. The first is that an independent EB would mitigate risks of manipulation. The relevant literature shows evidence of, and consensus on, the risk of manipulation of attempts to interpret ethical principles (Fazelpour & Lipton, 2020; Krishnan, 2020; Terzis, 2020). As Morley et al. stress:

> "AI practitioners may choose the translational tool that aligns with what is for them the most convenient epistemological understanding of an ethical principle, rather than the one that aligns with society's preferred understanding", (Morley et al., 2021, 243).

This creates risks of ethics shopping and ethics washing (Floridi, 2019), as well as reputational damage linked to these malpractices. The second reason concerns avoiding the risk of *ethics devolution*. This is a malpractice which occurs when the burden of interpreting AI ethics principles and for identifying criteria to implement the principles are shifted from institutions to their members (employees), who may lack the necessary expertise, resources, and understanding of ethical risks to make normative decisions, e.g. balancing conflicting principles. Ethics devolution can lead to oversimplification of ethical challenges, for example when these are reduced to health and safety issues, and trivialisation of ethical solutions and malpractices, like ethics blue-washing and ethics dumping (Floridi, 2019). Without the operational independence and justification provided by the multistakeholder EB, personnel could consider efforts to develop and apply ethical guidelines as a mere add-on or an extra burden, undermining both the development of a pro-ethical institutional culture and the outcome of any AI ethics initiative. Third, resulting from the first two reasons, an independent board could foster a genuine pro-ethical attitude within a defence organisation, whose members would perceive that the institutional focus on ethics is not superficial and that they can rely on independent expertise to identify and mitigate ethical risks.

For the EB to be effective, it is crucial that all relevant categories of stakeholders are involved, and that their involvement entails an active role in the shaping of the guidelines. At the very minimum this should include defence institutions and representatives of civil society (we take this category to represent the interests of noncombatants) as well as non-governmental organisations (e.g. the International Committee of the Red Cross -ICRC) to ensure a wider representation of noncombatant stakeholders, and technology providers.

Davies et al. (2015) identify two approaches for the involvement of stakeholders in the process of specifying guidelines: dialogical or consultative. They further distinguish.

> "*how* a normative conclusion can be justified, the *analytic process* through which that conclusion is reached, and the *kind* of conclusions that are sought. […] "the analytic process through which that conclusion is reached: should we prioritise the thinker, the theory, or the stakeholders? [… And]"the kind of

conclusion that is sought: should we aim for particularity of generalisability?" (Davies et al., 2015).

Under dialogical approaches, ethical analysis is part of the stakeholder involvement process itself. In this case, 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 the actual outcome or solution.

Under consultative approaches, ethical analysis is undertaken after an explicit engagement with stakeholders e.g., via a workshop, focus group, or deliberative mini-public. However, whilst stakeholder views feed into the ethical analysis, they are not involved in it. With consultative approaches, normative conclusions are justified on the basis of the coherence of the proposed solutions with the adopted moral theory, i.e. consistency with background principles and cases, in reflective equilibrium (Davies et al., 2015). The key element here is the definition of 'coherence'. For example, Dunn et al. (2012) define coherence in reference to argumentative standards in real world practical reasoning, rather than in terms of alignment with between moral theory and empirical facts. Here, justification is based on coherence between universal standards of justification and relevant contextual considerations. Others define 'coherence' following either a narrow or wide reflective equilibrium (RE) (Rawls, 1999). Narrow RE looks to find coherence between data and theory, and wide RE looks to find coherence among a number of relevant factors considered of equal importance, e.g. data, theory, moral principles, moral experiences of others, morally relevant facts and other background theories. Many of the broader consultative approaches in bioethics are based on wide RE. However, some reject that all factors have equal weight, for example approaches based on reflexive balancing stress that some factors are "quasi foundational moral principles" around which coherence must be built.

Both approaches offer important insights when considering the definition of ethical guidelines in the defence domain. For example, the discursive element of the dialogical approach and the need to develop consensus around ethical risks and desirable solutions is key when developing ethical guidelines for AI systems which will impact different stakeholder groups in different ways. At the same time, the debate on reflective equilibrium in consultative approaches is central when considering how to balance competing ethical principles in specific contexts.[5]

Adopting the distinction proposed by Davies and colleagues, we submit that the EB will work best if constructed following the dialogical approach. Provided

[5] Reflexive balancing is the process of arriving at ethical decisions through the balancing of competing relevant principles and contextual concerns, which may involve rejecting or amending existing beliefs or principles (Ives 2014).

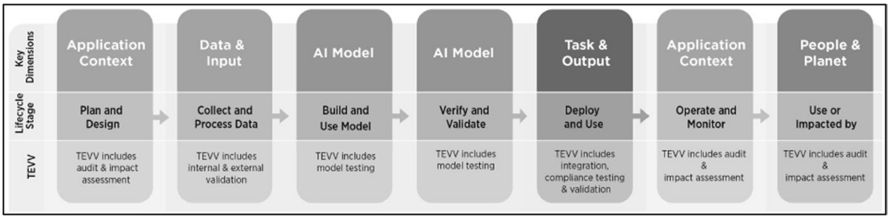
that there is an appropriate representation of all legitimate interests, and dialogue is conducted in accordance with appropriate background norms (e.g., transparency, reproducibility, military ethics, just war theory), democratic values, and within a pro-ethical institutional culture, consensus reached through dialogue, and active participation of stakeholders, will provide the solid normative justification of the decision of the EB. This is because, with adequate representation of the different stakeholders and appropriate ground rules, an EB designed to work according the dialogical approach can aim to achieve moral impartiality (Habermas, 1990) and to find fair ways to reconcile different interests (McCarthy, 1995; J. Heath, 2014), i.e. the board should produce recommendations with consequences that can be accepted as fair by all involved parties. This is crucial given the high-risk domain we are considering.

To achieve moral impartiality, the EB should work following Habermas' theory of discourse ethics (Habermas, 1990; Habermas et al., 2010; Habermas, 2021) whereby, through dialogue (discourse) the stakeholders commit to consider rationally the interests of the others, argue for their own positions, and work to find universally agreeable norms, i.e. to resolve a specific conflict of interests by identifying norms whose effect could be accepted as fair by all parties. This will be particularly relevant when the EB has to give indications as to how to balance ethical principles in specific contexts. We now focus on the first step of the methodology, abstraction.

2.2 Abstraction

The consensus in the relevant literature is that AI ethical guidelines should span the entire lifecycle of an AI system (Alshammari & Simpson, 2017; d'Aquin et al., 2018; Leslie, 2019; Department of Defence, 2022; Cihon et al., 2021; Dunnmon et al., 2021; High-Level Expert Group on Artificial Intelligence, 2019; Ayling & Chapman, 2022; Mäntymäki et al., 2022). This tallies with broader consensus that AI ethics governance must be holistic and systemic to be effective (Eitel-Porter, 2021). This is important from a process, product, and purpose point of view (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. From a product point of view, 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. From a purpose perspective, the social and political motivations of a project and the goals or trajectories of innovation may change over time, thus ensuring control over the project requires continuous monitoring of its ethical implications.

We agree that ethical guidelines need to address the entire lifecycle of the AI system. We submit that insofar as the AI lifecycle is a model of the processes and conditions of the design, development and deployment of AI technologies, it carries normative value, because those who define the AI lifecycle define the scope of applicability of both the principles and the guidelines. This is why the modelling of the AI lifecycle is a task for the EB.



Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment

Fig. 2 The NIST AI RMF Lifecycle (NIST, 2023, 11)

A key factor in modelling the AI lifecycle is the granularity of the model. A lifecycle with insufficiently differentiated stages can lead to blind spots and the creation of ethical risks. Yet, if too many stages (with related tasks) are identified, the iterative application of the principles multiplies, making the guidelines unwieldy. An approach that is too granular also risks being of little value, as it may be outdated by rapid developments in AI that alter the lifecycle stages. Help in finding the right granularity to model the AI lifecycle may be provided by existing standards and frameworks. In this case, however, issues concerning scope and purpose emerge.

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. Consider, for example, the National Institute of Standards & Technology (NIST) AI Risk Management Framework (RMF) (NIST, 2023). The framework is use-case agnostic, intended for organisations designing, developing, and deploying AI systems across different domains and sectors. Its purpose is to aid the identification and mitigation of risk, for which it includes an AI lifecycle model as well as the agents associated with each stage of the lifecycle and their responsibilities for managing AI risk (Fig. 2). The NIST model of the AI lifecycle improves on others, but since it was not designed to aid the interpretation of ethics principles it has severe limitations when used to this end. Without the flexibility to identify the salient steps in the AI lifecycle on a case-by-case basis, a context-agnostic framework runs the risk of pitching ethical analysis at a level which is at odds with case-specific requirements such as the relevant military development procedures, proportionality of the assessment, or simply missing key steps in the lifecycle where ethical risks emerge.

The choice as to how to model the AI lifecycle is not a trivial one. To facilitate this choice we use *Levels of Abstractions* (LoAs) (Floridi, 2008). Before proceeding, a brief introduction to the LoAs is required. LoAs are used in Systems Engineering and Computer Science to design models of a given system (Hoare, 1972; Heath et al., 1994; Diller, 1994; Jacky, 1997; Boca, 2014). They are also widely used in Digital Ethics (Floridi & Taddeo, 2016) and have been applied in this field of research to address several key issues, like identifying the responsibilities of online service providers (Taddeo & Floridi, 2015), offering guidance on the deployment of tracing and tracking technologies during the COVID 19 pandemic (Morley et al., 2020a, 2020b), analysing the possibilities of deterrence in cyberspace (Taddeo, 2017a), and considering the ethical implications of trust in digital technologies Taddeo, 2017b).

Any given system, for example a car, can be observed by focusing on specific properties while disregarding others. The choice of these aspects, i.e. the observables, depends on the observer's purpose or goal. An engineer interested in maximising the aerodynamics of a car may focus upon the shape of its parts, weight, and materials. A customer interested in the aesthetics of the car may focus on its colour and overall look. The engineer and the customer observe the same car at different LoAs. Thus, a LoA is a finite but non-empty set of observables accompanied by a statement of what feature of the system under consideration such a LoA stands for. A collection of LoAs constitutes an interface. An interface is used when analysing a system from various points of view, that is, at varying LoAs. It is important to stress that LoAs do not have to be hierarchical (though they can be): the engineer's and the user's LoAs are not one higher or lower than the other. And note that a single LoA does not reduce a car to merely the aerodynamics of its parts or to its overall look. Rather, a LoA is a tool that helps to make explicit the observation perspective and constrain it to only those elements that are relevant in a particular observation for the chosen purpose (Floridi, 2008).

When considering the AI lifecycle, the correct LoA is the one that allows for identifying the points in the lifecycle where the ethical risks that the principles tackle emerge, and the related accountable actors. This is why we chose a GoA (henceforth $\mathrm{GoA}_{\mathrm{Ethics}}$) that combines three LoAs focusing on: steps of the AI lifecycle ($\mathrm{LoA}_{\mathrm{Steps}}$), the actors accountable for those steps ($\mathrm{LoA}_{\mathrm{Actors}}$), and the ethical risks that can emerge at those steps ($\mathrm{LoA}_{\text{e-risks}}$).

As mentioned above, the granularity of the analysis is crucial for feasible, agile, and impactful guidelines. To avoid the risks related to too low or too high granularity, a solution is to focus on the steps of the AI lifecycle and disregard broader stages (too low granularity) and the specific tasks that a step entails (too high granularity). As for $\mathrm{LoA}_{\mathrm{actors}}$, we include both those who provide the technology or contribute to its design and development (we refer to them as to the provider) and those who decide on, act on, and monitor the deployment of an AI system (we refer to them as to the users). Both companies providing the technologies and defence organisations developing or using them have internal hierarchies and structures that shape the ways in which different parts of the organisations are involved in each step of the AI lifecycle. Thus, we suggest that actors accountable for the implementation of the guidelines for the different steps are identified following existing structures.

Below, we offer an example of how a model of the AI lifecycle would look given the proposed GoA. It is worth stressing here, that it is part of the tasks of the EB to identify the correct LoA and specify ethical risks accordingly. The example below is simply a way to outline a possible model. The proposed model results from adapting a model of the AI lifecycle proposed in (Floridi et al., 2022). We chose this model

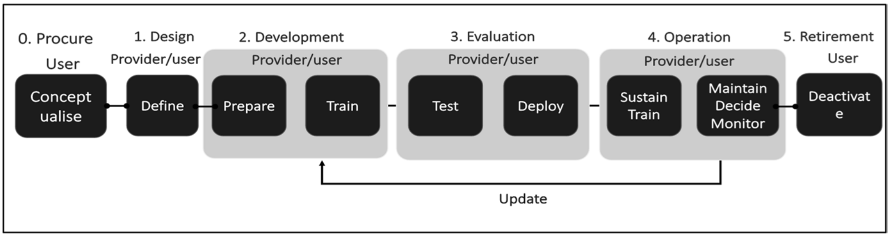


Fig. 3 The AI lifecycle modelled using the highly-granular GoA$_{\text{Ethics}}$. This model is adapted from (Floridi et al., 2022)

because it rests on widely adopted standards for software development such as ISO, IEC, and IEEE standards (ISO/IEC TR 24748–1[6] and ISO/IEC/IEEE 12207:2017)[7][8] and also because this model has been proposed to define a process to implement ethics-based auditing of the use of AI (Floridi et al., 2022). Thus, it already includes ethically relevant aspects. For example, it includes an evaluation stage, which is relevant when considering high-risks applications of AI such as those in the defence domain. The original model includes five stages – i.e. design, development, evaluation, operation, and retirement. We adapted it to include a procurement stage, as shown in Fig. 3, to identify those points in an AI lifecycle where decisions about developing new AI systems can lead to unethical consequences. Procurement (and conceptualisation) steps and design (and definition) steps are distinguished, where the former includes the specification of a possible use case and consideration of a system's utility and impact as well as requirements for providers; the latter refers to the definition of specific technical requirements, e.g. data needs and architecture of the AI model, that will inform the actual end system to be used.

Table 1 below shows the resulting model of the AI lifecycle using GoA$_{\text{Ethics}}$. The model sets out how ethical requirements resulting from the interpretation of the ethical principles should be defined to address the ethical risks related to the steps of an AI lifecycle model, and offers an indication as to which actors are responsible for meeting these requirements. The model could be expanded to consider specific steps related to the way a project is handled and deployed in a defence organisation. For example, an EB working in the UK could consider expanding the model to include specific steps of the CADMID cycle,[9] whenever these are deemed to foster specific ethics risks. The following section explores the methodology for interpreting ethical principles in more detail.

[6] https://www.iso.org/standard/72896.html

[7] https://www.iso.org/standard/63712.html

[8] It is worth noticing that standards to model the AI lifecycle are now emerging, see for example the ISO/IEC DIS5338 https://www.iso.org/standard/81118.html, an EB could start from here to design an ethics model of the lifecycle of an AI system.

[9] The CADMID cycle standards for concept, assessment, demonstration, manufacture, in-service, and disposal. It is the high-level project lifecycle used by all branches of the UK Ministry of Defence, and is particularly important for safety management. https://www.rpsgroup.com/cadmid-cycle/#:~:text=What%20is%20the%20CADMID%20cycle,and%20Disposal%20(CADMID)%20cycle.

Table 1 The model of the AI lifecycle at GoA$_{\text{Ethics}}$

Lifecycle stage	Accountable actor	Steps	Example of ethical risks
Procurement	User	• Conceptualize use case, context, architecture, and objective • Specify user requirements • Specify concept of employment • Assess fitness of the AI solution for the problem • Request for provider's qualifications	• Disproportionate solution • Lack of transparency of the AI model • Responsibility and accountability gap • Lack of transparency and traceability from the provider
Design	User/Provider	• Define data needs • Define trade-offs of algorithmic decisions • Provide risk analysis and define risk thresholds	• Limited robustness of the model • Responsibility gap • Disproportionate data collection (privacy breaches) • Transparency/efficiency balance inadequate • Design assumptions do not account adequately for contextual factors (e.g. racism, economic factors, complex environment)
Development	User/Provider	• Sourcing data • Data analysis • Preparing data • Splitting data • Build and Train an initial model • Develop a benchmark	• Data collected without proper consent • Model drift
Evaluation	User/Provider	• Test for undue outcomes, e.g. bias • Test for robustness • Evaluate primary metrics • Refine the model • Select deployment strategy	• Undue discrimination • Limited predictability • Accountability and responsibility gap • Vague specifications for deployment leading to unforeseen outcomes
Operation	User	• Monitor and trace • Post-deployment review • Define accountability • Establish feedback mechanism	• Accountability gap • Improper use leading to unforeseen outcomes • Communication flaws leading to accountability gaps
Retirement	User	• Assess deactivation risks • Archive logs	• Lack of logs

2.3 Interpretation and Requirements Elicitation

As noted in the introduction, AI ethics principles have been compared to constitutional principles. Like constitutional principles, ethical principles are meant to be foundational rather than offering detailed guidelines: they embed values more than specific directives, are expressed in simple and plain language, and they have an "open-textured character [… and a] purpose-oriented nature" (Dehousse, 1998, 76). They are often articulated in a non-hierarchical way, but they are competitive and may need to be balanced against each other depending on the specific context of application. These shared characteristics make judiciary methodologies to interpret constitutional principles effective in aiding the interpretation of ethics principles.

The relevant literature identifies five methodologies with which a constitutional judge may interpret constitutional principles (Llorens, 1999): literal, historical, contextual, comparative, and teleological.[10] For the interpretation of AI ethics principles, we refer to the teleological methodology and disregard the others.[11] This is because the literal and contextual methodologies look at the exact meaning of the words of the principles and their immediate context, respectively, to understand the prescription but disregarding the overall goal of the principles. Because of this, these two methodologies can lead to an interpretation inconsistent with the objective of the legislation in question (Llorens, 1999). Historical methodologies focus on the intention of the legislator and/or the function of the principles at the moment of their ratification. Here, the intentions of the legislator (a parliament) are considered insofar as the legislator is a representative body expressing the will of the public. This is not the case when considering the institutions that have drafted AI ethics principles. AI ethics principles are not formally ratified statutes and are not the product of legislators with express intent. Thus, this methodology is not fit for the purpose of our analysis. The comparative methodology considers the interpretation of similar principles adopted by other courts of justice. In the case of AI ethics principles, there is not yet an established tradition or approach to interpreting the principles, so this methodology is unfeasible.

The teleological methodology focuses on the purpose underpinning the principles, on their context and goal, and it is committed to reaching an *effet utile* (useful effect). It rests on Article 31.1 of The Vienna Convention on the Law of Treaties 1969, which states that.

[10] The literature on jurisprudence of the courts of justice is vast and methodologies for judiciary interpretation is a central topic of discussion, being the methods with which a judge assesses cases as a matter of justice administration, division of powers in democracies, and rule of law. We can disregard these issues when considering the interpretation of the ethical principles, which are voluntary principles that organisation adopt beyond legal compliance.

[11] The teleological interpretation of constitutional principles is not dissimilar from the *ratio legis* criterion (Canale and Tuzet 2010). The similarity with the teleological methodology is quite clear, the difference remains in reasons behind the application of the *ratio legis* and of the teleological methodology. Judges resort to the former to supplement the limited clarity of the text of the law, particularly civil laws; whereas the latter is one of the possible methodologies that a judge is *expected* to use to interpret constitutional principles, whose application requires an interpretation of some sort.

> "[a] treaty shall be interpreted in good faith in accordance with the ordinary meaning to be given to the terms of the Treaty in their context and in the light of its object and *purpose.*"[12]

The teleological interpretation looks at the words of the principles to identify their spirit, that is the essential values and rights that principles aim to protect. It considers context and objective, which refer to the context in which a principle is stated, for example the specific treaty, and goals of the treaty of which it is a part. Decisions about how to interpret a principle can be aided by considering the so-called *travaux préparatoires*, i.e. the available documents produced to aid the drafting of the principles themselves, which can be used to clarify their purpose. The methodology must be *effet utile*, that is, once the purpose of a principle is identified, it will be interpreted so to achieve effectiveness, consistency, and uniformity with the legal framework of a given state (Brittain, 2016; Fennelly, 1997).

Following the teleological methodology, the EB shall first consider the spirit of the principles to identify the values and rights that that they protect. Here, the discourse ethics theory described in Sect. 2 will be crucial to reconcile the views of the different stakeholders. The board should consider the context and the objective. In our example of the ethical principles issued by the Ministry of Defence, the context can be identified straightforwardly, for it refers to sets of values, e.g. democratic values, military ethics values, and ethical principles, such as those for example offered by Just War Theory (Blanchard & Taddeo, 2022).

The objective refers to the overall goal of the principles of identifying and mitigating ethical risks. Finally, respecting the commitment to ensuring useful outcomes from the principles, the board will have to define effective, applicable, measures to ensure that the objective is met and the values or rights protected in the principles are not breached. In view of the *effect util* element, the board should define requirements that need to be met at each step of the AI lifecycle in order to avoid breaching the AI ethics principles. In other words, the EB will work to answer the question 'what do I (provider/designer/developer/user) have to do to ensure that this step of the AI lifecycle respect the AI ethics principles?'. For example, when interpreting the principle of 'justified and overridable uses', the EB could specify the following requirements for an AI system to move from the procurement to the design step:

- There has to be an analysis mapping ethical risks related to the use of the envisaged AI system for a specified goal and detailing mitigation strategies for such risks;
- There has to be a cost–benefit analysis showing the military (or organizational) value of the proposed solution which demonstrates beneficial outcomes and proportionality of possible breaches of rights and values;
- The conceptualization of the AI solution should specify procedures to ensure human overridability of the system.

[12] Vienna Convention on the Law of Treaties, opened for signature, May 23, 1969, 1155 U.N.T.S. 331, 8 I.L.M. 340, 8 I.L.M. at 691–92.

The organisation receiving the recommendations of the board should consider the requirements to be necessary conditions for an AI system to move through its lifecycle and to ensure that the necessary measures are in place to mitigate the ethical risks identified by the EB. We do not focus on the operationalisation of these requirements and their verification, as these also depend on specific internal organisational policies, but it is worth stressing that efforts to translate or interpret ethical principles into practices become futile if, once translated, the guidelines are not adopted and respected at institutional level.

2.4 Balancing the Principles

The third and last step of the methodology concerns the balancing of the principles. There may be circumstances in which ethical principles have to be balanced against each other. One may imagine, for example, a case where one AI system is more robust but less transparent than an alternative option. In this respect ethical principles recall the constitutional ones, which often compete. In this case courts often establish a context-dependant hierarchy. As Alexy puts it:

> "the proposition of finding a "*conditional precedence relation*"(Alexy, 2002, 52): if the conditions x are given, [principle 1] prevails over [principles 2]; if the conditions y are given, [principle 2] prevails over [principle 1]." (emphasis added, (as reported in Guastini, 2019, 312)).

The definition of the conditional precedence relation is key to a fair balancing of the principles, but it can be tricky, as it has to consider institutional and cultural elements of an organisation as well as to ensure the resulting balancing leads to outcomes consistent with the overall ethos of the defence institution and with democratic values. Here, the independence and multistakeholder nature of the board offer assurances as to the fairness of the defined conditional precedence relations, while the teleological methodology offers guidance as to how to ponder and balance competing principles in specific cases. The balancing should be purpose- and context-dependant, that is it should focus on the ethical risks, particularly on the risk magnitude and impact of the AI lifecycle with respect to the purpose and use of the AI system – whether sustainment and support, adversarial and non-kinetic, and adversarial and kinetic – and the specific conditions of deployment. For example, the EB may have to consider the theatre of a specific kinetic operation. One way in which the EB could provide effective guidance is by expressing the conditional precedence relation in terms of purpose- and context-specific tolerance thresholds (i.e. how strictly a requirement needs to be met) of satisfaction of ethical requirements.

Thus, the balancing of the principles should include an assessment of the likelihood and of the impact of the ethical risks for specific purposes and contexts of use (Taddeo et al., 2021; Novelli et al., 2023). For example, lack of transparency and traceability on the provider's side for an AI system to be used for sustainment and support, and for an AI system to be used for kinetic and adversarial use may create the same risk of a responsibility gap. However, should the risk materialise, its impact would be much higher for kinetic and adversarial uses. This may guide the

EB to set a much lower tolerance threshold (leaving little flexibility) for the satisfaction of the ethical requirement for the kinetic and adversarial uses than for sustainment and support uses.

3 The Application of the Method to the Principles: Applying the Human-centric AI Principles in Procurement

In this section, we offer an example of the requirements that an EB may elicit when considering specific ethics principles for AI in defence. A clarification is necessary before introducing this example: the following simply describes the *type* of requirements and recommendations that the EB should provide. The measures indicated below are purely hypothetical as they were not developed by an EB like the one described in this article. Our example focuses on the application of the UK MoD AI principle of human-centricity to the procurement steps of the AI lifecycle. The principle states that.

> “The impact of AI-enabled systems on humans must be assessed and considered, for a full range of effects both positive and negative across the entire system lifecycle.
> Whether they are MOD personnel, civilians, or targets of military action, humans interacting with or affected by AI-enabled systems for Defence must be treated with respect. This means assessing and carefully considering the effects on humans of AI-enabled systems, taking full account of human diversity, and ensuring those effects are as positive as possible. These effects should prioritise human life and wellbeing, as well as wider concerns for human-kind such as environmental impacts, while taking account of military necessity. This applies across all uses of AI-enabled systems, from the back office to the battlefield.
> The choice to develop and deploy AI systems is an ethical one, which must be taken with human implications in mind. It may be unethical to use certain systems where negative human impacts outweigh the benefits. Conversely, there may be a strong ethical case for the development and use of an AI system where it would be demonstrably beneficial or result in a more ethical outcome” (Ministry of Defence, 2022, 9).

The EB would use the AI lifecycle model described in Table 1. and would interpret the principle to identify its spirit. In this case the spirit mandates respect for humans who may be subject to the actions of an AI system and for the environment, and to balance this respect with military necessity. Following the teleological methodology, the EB should consider the overall context and goal of the principle, and whether and how other principles informing the institutional culture of the MoD share the same spirit and what measures have been taken in the past to apply these principles. At the same time, as literature on the ethical risks of AI systems continues to flourish, the board should also rely on expertise, lessons learned, and existing good practices in defence and in other domains to draw ideas as to how to respect the spirit of this principle.

The interpretation of the principle with respect to step-specific ethical risks, as identified by the EB on the basis of the model of the AI lifecycle adopted, will lead the board to define the requirements as necessary steps for a project to move from the procurement stage to the next one. Depending on the institutional context, the board may be called upon where necessary to adjudicate on whether the requirements have been adequately fulfilled, to enable a project to move to the next lifecycle stage.

The requirements could be, for example:

- There has to be an analysis of the ethical risk of the envisaged AI system and an analysis showing an effective mitigation strategy
- There has to be a cost–benefit analysis showing the military (or organizational) value of the proposed solution which demonstrates beneficial outcomes and proportionality of possible breaches of rights and values
- There has to be a comparative sustainability analysis showing positive evidence of the sustainability of the envisaged solution, i.e. the proposed solution is more sustainable than alternatives
- There has to be evidence that the prospective provider has effective transparency policies, robust traceability protocols, and lack of negative records in terms of sharing necessary information when needed.

As discussed in Sect. 2.4, balancing is required to balance conflicting or competing principles and should include an assessment of the likelihood and of the impact of the ethical risks for specific purposes and contexts of use (Taddeo et al., 2021). Considering the requirements above, the EB has to also ensure that the balancing of principles and requirements leads to outcomes consistent with the overall ethos of the defence institution and with democratic values.

4 Conclusion

In this article, we have focused on the need to develop a methodology to interpret and apply AI ethics principle to specific practices, in this case those of the defence domain. To be effective, such a methodology has to produce applicable guidelines leading to fair solutions, which are consistent with the spirit of the AI ethics principles as well as with the foundational values of democratic societies. The methodology also has to be replicable and scrutinisable, so that it can be refined and adapted when its applications show flaws or limitations. With it, we aim to close the gap highlighted in Sect. 1 and provide the necessary middle step for the interpretation of AI ethics principles and their application into practices.

Questions concerning the operationalisation of the methodology are outside the scope of this article. However, we shall submit that the proposed methodology has great potential to be practical and agile, because following it an EB can identify ethical risks and requirements focusing on *types* of AI systems and for *purposes of use*. In this way, it is feasible that an EB will not have to consider and specify ethical requirements for every new AI system to be procured, developed, designed or used by a defence organisation. It may be the case that risks thresholds could be set so that under a certain threshold, a default set of ethical requirements provided by the EB are applied and above a certain threshold the EB will have to consider specific cases.

It has to be noted that the methodology is a necessary but insufficient element to foster ethical design, development and deployment of AI systems. To achieve this result, and to avoid risks of malpractice, these efforts need to stem from a pro-ethical institutional culture, where ethics is not perceived or treated as an add-on or an extra burden for practitioners, but is a constitutive, unavoidable element of everyday practices, which aids the achievement of positive results. This is especially the case when considering defence institutions or other public bodies working in high-risk domains.

Such a pro-ethical attitude needs to be fostered and demonstrated at an institutional level. Two ways are particularly important: ethics training and enforceability. While we have argued in this article that ethical decisions are too complex to be devolved to individual practitioners or groups of practitioners without the required expertise, ethical outcomes are fostered when practitioners are aware of ethical risks, problems, complexities, and opportunities coupled to the use of AI, and which make necessary the adherence to the specified ethical guidelines. Thus, ethics training should be provided at the institutional level and made accessible (if not mandatory) to practitioners. At the same time, ethical guidelines that are not enforced would undermine any efforts to develop ethical uses of AI. To this end it is crucial that accountability is clearly established, for example via an EBA process.

We leave these topics to further steps in our research; but it is crucial to mention them here to outline, albeit briefly, that a methodology without institutional support would not be more useful to achieve ethical outcomes than the AI ethics principles alone.

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 David McNeish and James Rosie for helpful comments on the draft.

Author's Contributions MT, AB, TC.

Funding Mariarosaria Taddeo, Alexander Blanchard, and Christopher Thomas 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 NA, there are no data included in this study.

Declarations

Ethics Approval and Consent to Participate Not needed, no data was collected for this study.

Consent for Publication Not needed, no data was collected for this study.

Competing Interests Mariarosaria Taddeo is a member of the Ethics Advisory Board of the UK Ministry of Defence.

References

Alexy, R. (2002). *A Theory of Constitutional Rights*. Oxford University Press.

Alshammari, M., & Simpson, A. (2017). Towards a Principled Approach for Engineering Privacy by Design. In E. Schweighofer, H. Leitold, A. Mitrakas & K. Rannenberg (Eds.), *Privacy Technologies and Policy. Lecture Notes in Computer Science* (10518, 161–77). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-67280-9_9

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias'. *ProPublica*, 23 May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aquin, M. D'., Troullinou, P., O'Connor, N. E., 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

Blanchard, A., & Taddeo, M. (2022). Autonomous weapon systems and jus Ad bellum'. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01425-y

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

Blanchard, A., Thomas, C., & Taddeo, M. (2024). Ethical governance of artificial intelligence for defence: *Normative Tradeoffs for Principle to Practice Guidance'. AI and Society*, Springer (fothcoming).

Boca, P. (2014). *Formal Methods: State of the Art and New Directions*. London: Springer.

Brittain, S. (2016). Justifying the Teleological Methodology of the European Court of Justice: A Rebuttal. *Irish Jurist, New Series, 55*, 134–165.

Canale, D., & Tuzet, G. (2010). What Is the Reason for This Rule? An Inferential Account of the Ratio Legis. *Argumentation, 24*(2), 197–210. https://doi.org/10.1007/s10503-009-9171-x

Cihon, P., Schuett, J., & Baum, S. D. (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

Davies, R., Ives, J., & Dunn, M. (2015). A Systematic Review of Empirical Bioethics Methodologies. *BMC Medical Ethics, 16*(1), 15. https://doi.org/10.1186/s12910-015-0010-3

Dehousse, R. (1998). *The European Court of Justice: The Politics of Judicial Integration. The European Union Series*. New York: St. Martin's Press.

Department of Defense. (2022). 'Responsible Artificial Intelligence Strategy and Implementation Pathway'. Virginia, United States: Department of Defense.

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

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

Diller, A. (1994). *Z: An Introduction to Formal Methods*. (2nd ed.) Chichester, West Sussex, England ; New York: Wiley & Sons.

Dunn, M., Sheehan, M., Hope, T., & Parker, M. (2012). Toward Methodological Innovation in Empirical Ethics Research. *Cambridge Quarterly of Healthcare Ethics, 21*(4), 466–480. https://doi.org/10.1017/S0963180112000242

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

Eitel-Porter, R. (2021). Beyond the Promise: Implementing Ethical AI. *AI and Ethics, 1*(1), 73–80. https://doi.org/10.1007/s43681-020-00011-6

Fazelpour, S., & Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, (pp 57–63). https://doi.org/10.1145/3375627.3375828

Fennelly, N. (1997). Legal Interpretation at the European Court of Justice. *FordhamInternationalLaw-Journal, 20*, 656–679.

Floridi, L. (2008). The Method of Levels of Abstraction. *Minds and Machines, 18*(3), 303–329. https://doi.org/10.1007/s11023-008-9113-7

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology, 32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi, L., & Taddeo, M. (2016). What Is Data Ethics? *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Floridi, L., Holweg, M., Taddeo, M., Silva, J. A., Mökander, J., & Wen, Y. (2022). capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064091

Georgieva, I., Lazo, C., Timan, T., & van Veenstra, A. F. (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

Guastini, R.(2019). *Identificazione, interpretazione dei principi costituzionali*. In . Rome: Università degli Studi di Roma 3.

Habermas, J., Cronin, C., & De Greiff, P. (2010). *The Inclusion of the Other: Studies in Political Theory*. Princeton.

Habermas, J. (1990). Habermas, Jürgen 1990. "Discourse Ethics, Notes on a Program of Philosophical Justification," in C. Lenhardt and S. W. Nicholsen (Trans.), Moral Consciousness and Communicative Action. Cambridge, MA: MIT Press, Pp. 43–115. In *Moral Consciousness and Communicative Action*, translated by C. Lenhardt and S. W. Nicholsen, 43–115. Cambridge, MA: MIT Press.

Habermas, J. (2021). *The Structural Transformation of the Public Sphere: ‡an ‡inquiry into a Category of Bourgeois Society*. Translated by Thomas Burger and Frederick G. Lawrence. Reprinted. Cambridge: Polity Press.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Heath, J. (2014). Rebooting Discourse Ethics. *Philosophy & Social Criticism, 40*(9), 829–866. https://doi.org/10.1177/0191453714545340

Heath, D., Allum, D., & Dunckley, L. (1994). *Introductory Logic and Formal Methods*. Henley-on-Thames: Alfred Waller.

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

Hoare, C. A. R. (1972). Structured Programming. In , edited by O. J. Dahl, E. W. Dijkstra, and C. A. R. Hoare, 83–174. London, UK, UK: Academic Press Ltd. http://dl.acm.org/citation.cfm?id=1243380.1243382

Ives, J. (2014). A Method of Reflexive Balancing in a Pragmatic, Interdisciplinary and Reflexive Bioethics. *Bioethics, 28*(6), 302–312. https://doi.org/10.1111/bioe.12018

Jacky, J. (1997). The *Way of Z: Practical Programming with Formal Methods*. Cambridge: Cambridge University Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Kim, S. Y. H., Wall, I. F., Stanczyk, A., & De Vries, R. (2009). Assessing the Public's Views in Research Ethics Controversies: Deliberative Democracy and Bioethics as Natural Allies. *Journal of Empirical Research on Human Research Ethics, 4*(4), 3–16. https://doi.org/10.1525/jer.2009.4.4.3

Krishnan, M. (2020). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology, 33*(3), 487–502. https://doi.org/10.1007/s13347-019-00372-9

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A Guide for the Responsible Design and Implementation of AI the Public Sector'. London: *The Alan Turing Institute*. https://doi.org/10.5281/ZENODO.3240529

Llorens, A. A. (1999). The European Court of Justice, More than a Teleological Court. *Cambridge Yearbook of European Legal Studies, 2*, 373–398. https://doi.org/10.5235/152888712802815789

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

McCarthy, T. (1995). Practical Discourse: On the Relation of Morality to Politics. *Revue Internationale De Philosophie, 49*(194), 461–481.

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

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

Morley, J., Cowls, J., Taddeo, M., & Floridi, L. (2020a). Ethical Guidelines for COVID-19 Tracing Apps. *Nature, 582*, 29–31.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020b). 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, 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. *Minds and Machines, 31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w

Munn, L. (2022). The Uselessness of AI Ethics. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00209-w

NIST. (2023). AI Risk Management Framework (AI RMF 1.0). NIST AI 100–1. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking AI Risks Seriously: A Proposal for the AI Act. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4447964

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

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

Rawls, J. (1999). *A Theory of Justice* (Revised). Belknap Press.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a Framework for Responsible Innovation. *Research Policy, 42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Taddeo, M. (2013). Cyber Security and Individual Rights, Striking the Right Balance. *Philosophy & Technology, 26*(4), 353–356. https://doi.org/10.1007/s13347-013-0140-9

Taddeo, M. (2014). The Struggle Between Liberties and Authorities in the Information Age. *Science and Engineering Ethics* pp.1–14. https://doi.org/10.1007/s11948-014-9586-0

Taddeo, M. (2017a). The Limits of Deterrence Theory in Cyberspace. *Philosophy & Technology*. https://doi.org/10.1007/s13347-017-0290-2

Taddeo, M. (2017b). Trusting Digital Technologies Correctly. *Minds and Machines 27*(4), 565–68. https://doi.org/10.1007/10.s11023-017-9450-5

Taddeo, M., & Floridi, L. (2015). The Debate on the Moral Responsibilities of Online Service Providers. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-015-9734-1

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science 361*(6404), 751–52. https://doi.org/10.1126/science.aat5991

Taddeo, M., McNeish, D., Blanchard A, & Edgar, E. (2021). Ethical principles for artificial intelligence in national defence. *Philosophy & Technology 34*(4):1707–29. https://doi.org/10.1007/s13347-021-00482-3

Taddeo, M., Ziosi, M., Tsamados, A., Gilli, L,. & Kurapati, S. (2022). Artificial Intelligence for National Security: The Predictability Problem. London: Centrefor Emerging Technology and Security.

Terzis, P. (2020). Onward for the freedom of others: Marching beyond the AI Ethics. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (pp. 220–229). https://doi.org/10.1145/3351095.3373152

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*, (195–200). 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