Resolving Ethics Trade-offs in Implementing Responsible AI

Conrad Sanderson [†◇], Emma Schleiger [†], David Douglas [‡], Petra Kuhnert [†], Qinghua Lu [†]

[†] *Data61 / CSIRO, Australia;* [‡] *Environment / CSIRO, Australia;* [◇] *Griffith University, Australia*

***Abstract*—While the operationalisation of high-level AI ethics principles into practical AI/ML systems has made progress, there is still a theory-practice gap in managing tensions between the underlying AI ethics aspects. We cover five approaches for addressing the tensions via trade-offs, ranging from rudimentary to complex. The approaches differ in the types of considered context, scope, methods for measuring contexts, and degree of justification. None of the approaches is likely to be appropriate for all organisations, systems, or applications. To address this, we propose a framework which consists of: (i) proactive identification of tensions, (ii) prioritisation and weighting of ethics aspects, (iii) justification and documentation of trade-off decisions. The proposed framework aims to facilitate the implementation of well-rounded AI/ML systems that are appropriate for potential regulatory requirements.**

***Index Terms*—AI ethics, responsible AI, tensions, trade-offs, system design, societal impact, governance, regulations.**

I. Introduction

The increasing impact of artificial intelligence (AI) and machine learning (ML) across many sectors of society has led to a broad consensus on the need to design and implement these technologies in a responsible manner. Globally, governments, organisations and industry groups have defined many sets of high-level AI ethics principles to support this vision [1]; an example is shown in Table I. Such high-level principles contain a set of underlying themes and aspects, which typically includes accuracy/performance, robustness/safety, fairness, privacy, explainability/interpretability, transparency, and accountability [1], [2].

As the high-level principles are implemented in practice, the underlying aspects interact, which inevitably leads to various tensions and trade-offs [3], [4], [5]. Among the many observed interactions (summarised in Table II), a prominent example is the trade-off between accuracy and explainability [6], [7], [8]. While the need to formally address the trade-offs is often stated in the literature [5], [9], [10], [11], there is currently no generally agreed upon framework to accomplish this.

A further complication is that many designers and developers of AI/ML systems[1] are currently unaware of the tensions and trade-offs, which may stem from unfamiliarity of (or their unwillingness to engage with) AI ethics principles and/or their underlying aspects [12], [13]. Without regulatory enforcement, taking AI ethics principles into account can be contrary to industry priorities [14]. For example, it has been observed that taking into account the fairness aspect can considerably reduce the accuracy of AI/ML systems, affecting the potential profitability to be gained from using these systems [15].

A recent analysis of highly cited research papers within AI/ML fields shows that they contain many potentially harmful implicit biases and assumptions, as well as an inherent selection and prioritisation of ethics aspects [16]. The accuracy aspect (indirectly represented as performance) is the most emphasised, at the cost of considerably de-emphasising almost all other aspects. The next most commonly prioritised quality is generalisation, which is haphazardly and inconsistently used as a proxy for the robustness aspect in AI/ML literature [17].

The selection, prioritisation and trade-off resolution of AI ethics aspects can occur at various points in the AI/ML system development pipeline. Without organisational policies and formal governance, these can occur on an ad-hoc basis at the design and implementation levels, and as such can be significantly affected by individual team members, their knowledge and interpretation of Responsible AI issues, personal preferences and bias [13], [18], and lack of understanding of the effect of trade-offs on others [19]. On the other hand, explicit organisational policies may be under-developed and/or can lead to overly rigid adherence due to lack of flexibility [13], [20].

We can consider a hierarchy of needs at three levels for informing the implementation of Responsible AI systems:

- **Societal level**. Governments, industry bodies and regulators determine high-level principles, standards and regulations for AI/ML system development and use [21], [22]. These are influenced by cultural norms, values and existing legislation.
- **Organisational level**. Instead of approaching trade-offs ad-hoc, acknowledge that a set of trade-offs exists [3]; ensure designers and developers are aware of these trade-offs; create frameworks and procedures for dealing with trade-offs that are aligned with organisational, societal and regulatory expectations [20].
- **Practitioner level**. Prevent personal bias going into trade-off decisions; be aware that there is more than one point of view [18] and that there may be implications for various groups [19]. Developers of AI/ML systems action accepted frameworks like risk assessments, justifications and patterns [23], in order to translate principles into practice.

Given the above considerations, an important next step is hence the development of frameworks and/or guidelines to provide approaches to manage tensions and resolve trade-offs between AI ethics aspects, in order to facilitate the design and implementation of well-rounded Responsible AI systems.

To that end, we summarise and analyse various approaches to addressing trade-offs in Sec. II, noting their advantages and disadvantages. We discuss the overall properties of the examined approaches in Sec. III, and propose a multi-step framework that draws on the gained insights as well as the needs at societal, organisational and practitioner levels. Concluding remarks are given in Sec. IV.

[1]In this work we use the term *AI/ML systems* to refer to both ML models (algorithms) and AI products.

979-8-3503-5409-6/24/$31.00 ©2024 Crown
DOI 10.1109/CAI59869.2024.00215

Table I: Summary of the high-level AI ethics principles proposed by the Australian Government [21].

Principle	Summary
Human, societal & environmental wellbeing	AI systems should benefit individuals, society, and the environment.
Human-centred Values	AI systems should respect human rights, diversity, and autonomy of individuals.
Fairness	AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.
Privacy Protection & Security	AI systems should ensure the security of data, as well as respecting and upholding privacy rights.
Reliability & Safety	AI systems should reliably operate in the context of their intended purpose.
Transparency & Explainability	There should be transparency and responsible disclosure so that people can determine that an AI system is engaging with them and may have significant impact on them. Explainability provides descriptions of what the AI system is doing and why, and may include the system's processes and input data.
Contestability	When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow challenging the use and outcomes of the AI system.
Accountability	Human oversight of AI systems should be enabled, and the people responsible for the various phases of the AI system lifecycle should be identifiable and accountable for the system's outcomes.

II. Approaches for Resolving Trade-offs

A. *Dominant Aspects*

A blunt and straightforward approach to resolve tensions between ethics aspects is to select the most dominant or pertinent aspect in a given context. The prioritisation of aspects (eg. accuracy over privacy) can be driven by how difficult or costly it is to implement a given aspect within an AI/ML system, and/or internal organisational needs (eg. regulatory compliance), and/or the preferences of end users of the AI/ML systems [18].

The advantage of this approach is its overall simplicity and the low degree of required effort. A major disadvantage is that this is a winner-takes-all approach, which leaves no room to devise balanced trade-offs and take into account nuance, which in turn implies that a thorough evaluation of associated risks and benefits is not performed. This approach is hence consistent with the pejorative notions of *ethical lip service* [14] and *ethics washing* [24], where only minimal effort is expended to address ethical issues that emerge in AI/ML systems.

Table II: Matrix of observed interactions between AI ethics aspects, summarised from [3]. ● = trade-off; ○ = synergistic interaction; ◐ = mixed or context dependent interaction.

	accuracy	robustness	fairness	privacy	explainability	transparency
accuracy		◐	●	●	●	
robustness	◐		◐	○		◐
fairness	●	◐		●		◐
privacy	●	○	●			◐
explainability	●					○
transparency		◐	◐	◐	○	

B. *Risk Reduction via Aspect Infringement and Amelioration*

An indirect approach to resolving trade-offs is via prioritisation of ethics aspects, as proposed in [25]. This involves a multi-step strategy to reduce the operational risk of AI/ML systems, summarised as follows. First, an undesirable operational event in a given AI/ML system is identified (eg. a specific failure), through a risk assessment matrix that takes into account the likelihood of the event and associated degree of loss. Secondly, the risk assessment matrix is expanded to allow degrees of infringement (de-prioritisation) of given ethics aspects in order to reduce the risk of undesirable events. Lastly, additional safeguards are put into place with the intention to ameliorate the infringement of the affected ethics aspects.

As per the example given in [25], an AI-based user authentication system (eg. to prevent unauthorised access to bank accounts) can be made more accurate and/or more robust (eg. less susceptible to impersonation attacks) by requiring the use of more personally identifiable information (eg. face images [26]). However, the use of such information "infringes" the privacy aspect. The infringement is then ameliorated through further security measures to protect the information and to comply with applicable laws such as EU's General Data Protection Regulation (GDPR) [27].

An advantage of the above multi-step strategy is that the prioritisation of ethics aspects is driven by system-specific requirements and takes into account the context of system operation. The disadvantage is that the risk assessment step may be error-prone (eg. missing undesirable events, unreliable estimation of likelihood and loss), and may require expert knowledge which is beyond the level available to the design and development teams. Furthermore, it may not be possible to adequately ameliorate the infringement of applicable ethics aspects. Lastly, the incorporation of ethics aspects is done during later stages of AI/ML system design, which can be interpreted as treating the aspects as add-ons, rather than taking them into account from the very outset.

C. Trade-Off Analysis in Requirements Engineering

Trade-off analysis techniques used in *requirements engineering* [29], [30], [31] may be applicable for determining how each ethics aspect affects an AI system under construction and for addressing the interplay between ethics aspects [28], [32]. Each applicable ethics aspect is listed, followed by listing possible ML models and data types that may be suitable for implementing an AI system. Linkages between the ethics aspects and system components (model types and data types) are then graphically noted, in conjunction with their positive or negative effect (either in a quantitative or qualitative manner).

An illustrative example of the above trade-off analysis approach is shown in Fig. 1. The AI system under construction is a biometric user authentication system employing speech and/or face data, with the overall goal to increase security of banking services. Accuracy of the AI system is driven by two main components: model type and data type [33]. Two ML models are considered: Hidden Markov Model (HMM) [26] and Deep Neural Network (DNN) [34]. Furthermore, two data types are considered: speech only, and speech in conjunction with face data. In this example, using DNN can increase the accuracy by 5%, but at the cost of reduced explainability in comparison to HMM. Using speech and face data over using speech data alone can increase accuracy by 10%, though this negatively affects the privacy aspect as more personally identifiable information is used.

This type of trade-off analysis techniques can be considered as part of the design phase, where many possible implementations of an overall AI/ML system are explored. The techniques can also be considered as part of the documentation phase, where the trade-offs are explicitly documented (rather than left as tacit knowledge), along with discussions on the practical pros and cons of each possible implementation. This ties in with the dimension of justification suggested in [4], where the justification provides a context-specific rationale for the drivers behind giving more weight to one aspect over another.

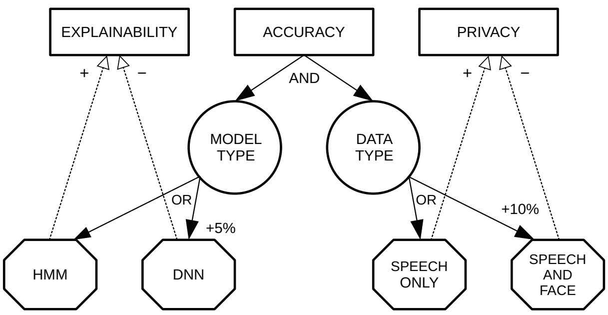

Figure 1: An example of graphical trade-off analysis, adapted from [28]. For an AI based authentication system using biometrics, the consideration involves two ML models (HMM and DNN) and two data types (speech only, and speech in conjunction with face).

An advantage of this approach is that many possible trade-offs can be explicitly shown and considered at the same time, and the consideration is done as part of the design phase. A disadvantage is that the graphical representation can become quite complex when more model types, data types and ethics aspects are considered. Furthermore, multiple graphs may be required, depending on the complexity of the ML pipeline [35].

D. Quantitative Ranking of Trade-Off Solutions

Inspired by [36], a ranking approach can be used to choose trade-off solutions. Given several technical solutions to a given trade-off, such as a set of possible ML models with various degrees of explainability for the accuracy/explainability trade-off [6], each solution is ranked according to an overall score.

The overall score is a weighted convex linear combination of a set of normalised sub-scores [37], with each sub-score representing how well a given solution addresses a desired characteristic pertinent to the trade-off at hand. The set of characteristics can range from purely pragmatic (eg. complexity of the ML model), to various philosophical positions (such as utilitarianism and egalitarianism [36]). The weighting of the characteristics can be non-informative (ie. all weights are equal and sum to one), or it can be based on the importance of each characteristic to an organisation (ie. weights are unequal and skewed towards focusing on a subset of characteristics).

The main advantage of this approach is that it aims to provide a quantitative procedure for resolving trade-offs, and explicitly allows for the consideration of characteristics that are important for the technical implementation of the AI/ML system, as well as wider organisational policies. However, the flexibility can also be a disadvantage, in that the selection of characteristics can be subject to errors, and hence may require well-informed reasoning that may be beyond the capability of the practitioners and/or the organisation. Furthermore, determining sub-scores and associated weights for the characteristics can be subjective, especially when dealing with characteristics that are not easily quantifiable [38]. For example, it is non-trivial to represent the degree of egalitarianism as a precise numeric value.

E. Specification and Balancing via Principlism

Principlism is an influential approach in bioethics that uses a set of moral principles to guide ethical decisions, such as respect for autonomy, nonmaleficence (avoid causing harm), beneficence (promoting the welfare of others), and justice [39]. The principlism approach has also been applied to cybersecurity ethics [40]. The similarity with high-level AI ethics principles makes principlism a useful approach to draw on when considering how to address trade-offs between the underlying AI ethics aspects, and for identifying the limitations of using a principlist approach to AI ethics [41], [42].

Principlism uses two approaches to bridge the gap between abstract principles and addressing individual cases [39]: (i) specification, and (ii) balancing. Specification elaborates on the principles to describe how individual cases are relevant to a specific principle or to the underlying ethical aspect behind it. For example, the principle of justice may be further specified by a rule that prohibits using ethnicity or gender as a basis for distributing access to resources [43]. AI ethics principles may also come with brief descriptions of how each principle may be applied [21]. However, such elaborations of basic principles will not cover all the possible cases, and will not remove all the potential conflicts between them [39].

When ethics principles or their underlying aspects give conflicting recommendations, the principlism approach provides six conditions that any balancing or trade-off must meet [39], as summarised below:

1. A stronger justification can be given for prioritising one aspect over another.
2. The purpose of overriding a given aspect has a realistic chance of being achieved.
3. There are no alternatives to this trade-off that are morally preferable.
4. The overridden aspect is infringed to the smallest extent possible to achieve the purpose of overriding it.
5. The negative effects of overriding the aspect are minimised.
6. Those affected by this trade-off are treated impartially.

The justifications for prioritising a given aspect over another (condition 1) may be grounded in practical limitations or in ethical theory. Such justifications must still comply with the other five conditions. Ethical concepts or theories that may ground such trade-offs include:

- Proportionality: the methods used should be appropriate for the problem the AI/ML system is intended to solve, and should have a minimal impact on the system's compliance with the other aspects [44].
- Benefit to the least-advantaged (maximin): decisions should be made based on maximising the benefits to the worst-off [45], [46].
- Utilitarianism/Consequentialism: decisions should be made based on maximising the utility or total benefits to all those affected by them [46].

There are several advantages to adopting a principlist approach to addressing AI ethics concerns. A set of ethics principles offers developers a common vocabulary for discussing ethics issues [42]. Each principle serves as a starting point for further ethical reflection and discussion about how it may apply to a specific AI application. The explicit statement of principles can also encourage cultural shifts within professions towards prioritising the values and goals that they express [42]. They may serve to establish ethical norms and change the values that are prioritised within professional communities [42].

However, the principlism approach also has several disadvantages. While the six conditions described above provide guidelines for how to make trade-offs between ethics aspects, how developers respond to these conditions is left up to the developers' judgement [47]. One set of developers may take a consequentialist approach to justifying the trade-offs they make, while another may favour using maximin. As a result, various developers may make different trade-offs when faced with the same conflict between ethics aspects.

Furthermore, AI/ML system development occurs in a significantly different context to medicine, where principlism has been influential as an approach to ethics [41]. Unlike medicine, AI development does not have common goals, a long professional history that has clear descriptions of the ethical duties expected of practitioners, a relative lack of methods for interpreting principles into practice, and the relative lack of professional accountability [41]. These differences may limit the effectiveness of ethics principles for influencing AI design and governance [41].

III. Discussion and Recommendations

In the preceding section we examined five approaches that can be incorporated into the AI/ML system design process with the aim of addressing tensions between common AI ethics aspects. From across the covered approaches we can extract five properties: (1) prioritisation of ethics aspects based on risk assessment/context specific assessment; (2) proactive analysis at start of design process; (3) de-prioritisation of an ethics aspect in favour of another; (4) analysis requires deliberation and/or resources to enact; (5) additional considerations required to ameliorate side-effects. Table III indicates the presence and absence of the above properties for each of the approaches.

There is no single approach that is likely to be appropriate for all organisations, AI/ML systems, or applications. In order to address this shortcoming, we propose a layered framework to introduce into the AI/ML system design pipeline, which is enacted at a practitioner level while being supported by organisational and societal influence.

Guidance of the design process and high-level decision making at a *societal level* can be informed by scientific literature, AI ethics principles, standards and regulations. These dimensions, however, do not take into account context, which is a key consideration in design decisions when managing tensions. Context includes the purpose of the AI/ML system, the groups of people impacted by the system, and the level of understanding about the outputs of the system that are required [8]. At an *organisational level*, policies and governance can be more specific to the general context of an AI application [48], but will generally be agnostic to the nuanced details required to make design decisions on a case-by-case basis.

Table III: Properties of approaches in Sec. II for addressing trade-offs. (1) = prioritisation of ethics aspects based on risk assessment/context specific assessment; (2) = proactive analysis at start of design process; (3) = de-prioritisation of an ethics aspect in favour of another; (4) = requires deliberation/resources to enact; (5) = additional considerations required to ameliorate side-effects. ● = property present; ○ = property absent.

Approach	Property (1)	(2)	(3)	(4)	(5)
Dominant Aspects (Sec. II-A)	○	○	●	○	○
Risk Reduction (Sec. II-B)	●	○	●	●	●
Requirements Eng. (Sec. II-C)	●	●	●	●	○
Quantitative Ranking (Sec. II-D)	●	○	●	●	○
Principlism (Sec. II-E)	●	○	●	●	●

At the *practitioner level*, where the resolution of tensions is specific to the context of the AI model, risk assessments as well as design tools and justifications can be used.

We can use insights from the approaches described in Sec. II to build a multi-step framework around managing trade-offs and tensions between ethics aspects. The proposed framework is comprised of three main components: (i) proactive identification, (ii) prioritisation and weighting, (iii) justification and documentation. The components are elucidated below.

Proactive identification. A proactive and structured assessment at the beginning of the AI/ML design pipeline provides a framework to examine and define the context and purpose of the AI/ML system. It also encourages a proactive approach for the identification of potential tensions between ethics aspects as well as the consideration of the methodology required to resolve them. The *requirements engineering* approach (Sec. II-C) encompasses proactive design consideration where ethics aspects, models and data types are examined, and the associated tensions and solutions are explored [28]. Similarly, a Value Sensitive Design (VSD) approach exemplifies a proactive approach to applying high-level ethics principles in practice [4]. VSD is an iterative and multidisciplinary process where ethical considerations are incorporated into the design process from the start [49]. Other approaches consider tensions between ethics aspects as they arise in the design process. For example, ranking of trade-off solutions (Sec. II-D) would routinely occur when tensions arise and can be used in addition to a prospective risk analysis. Specifying what the ethics principles mean in the context of a specific AI/ML project (such as in the *principlism* approach in Sec. II-E) may also identify possible tensions between ethics aspects early in the design process and hence allow them to be addressed.

Prioritisation and weighting. The nature of resolving tensions between ethics aspects that arise when developing AI/ML systems is that trade-offs must be made, resulting in (a) one or more ethics aspects being prioritised over others, and/or (b) a weighted (balanced) combination of selected ethics aspects. The least critical technique to apply is the *dominant aspects* approach (Sec. II-A), where prioritisation is made based on one-dimensional decisions such as difficulty or cost. In contrast, the other approaches listed in Sec. II require prioritisation based on a degree of context-based assessment. To this end, the first step in prioritising one ethics aspect over another is a consideration of the broader context. This can be done in a qualitative manner to selectively introduce trade-offs in order to reduce identified operational risks, as per the *risk reduction* approach (Sec. II-B). Alternatively, the *quantitative ranking* approach (Sec. II-D) can be used, where various solutions are ranked according to desired characteristics. Such rankings can also inform the responses to the six conditions specified for resolving conflicts between ethics aspects used by the *principlism* approach (Sec. II-E). A hybrid approach is used by the *requirements engineering* approach (Sec. II-C), where both quantitative and qualitative measurements are employed.

Justification and documentation. Explicability and transparency are critical aspects of responsible design [8]. As part of that, documentation and reasoning are essential concerning the design decisions where trade-offs have been made. Justification provides a context-specific rationale for the drivers behind giving more weight to one ethics aspect over another [4]. The *requirements engineering* approach (Sec. II-C) can provide justification focused on practical aspects of AI/ML system functionality, while the *quantitative ranking* approach (Sec. II-D) can take into account dimensions that go beyond immediate practical aspects. The *principlism* approach (Sec. II-E) takes a further step and provides a strong framework to develop justifications, where prescriptive conditions are explored in the process and can then be used to document and justify trade-off decisions. For example, documentation that addresses how the six conditions described in Sec. II-E are being met through considered design decisions, so that there is accountability for how the design decisions were made. These explanations (and the impacts of the resulting trade-offs) may also inform future decisions about resolving ethical tensions that arise in other projects.

IV. Concluding Remarks

While progress has been made in the application of high-level AI ethics principles in the design and implementation of AI/ML systems [50], [51], [52], there are still notable areas of concern, including a theory-practice gap in how to manage tensions that arise between commonly accepted AI ethics aspects that underpin the principles [3], [4].

In this work, we have covered five approaches for addressing the tensions via trade-offs, ranging from rudimentary to quite complex. The approaches mainly differ in the types of considered context, the scope of each context, the associated methods for qualitatively and/or quantitatively measuring each context, and how each trade-off decision is justified. None of these approaches is likely to be appropriate for all organisations, AI/ML systems, or applications.

In response, we have proposed a framework for AI/ML system developers for use in the design pipeline that draws on the various strengths of the covered approaches. The framework has three main components: (i) proactive identification, (ii) prioritisation and weighting, (iii) justification and documentation. At the start of the AI/ML design pipeline, potential tensions between ethics aspects are identified, and consideration is given to the methodology for resolving them. The tensions between the identified aspects are then addressed through prioritisation and weighting, using one or more of the covered approaches; both practical and organisational requirements can be taken into account. Each trade-off decision is justified and documented, aiding transparency and accountability, as well as adding to the pool of organisational knowledge.

AI/ML systems built with rudimentary and/or shallow ethical assessments are unlikely to be robust against potential legal and regulatory challenges. Employing a proactive and dynamic assessment method across the full AI/ML system pipeline (including the context of the system's application), such as the framework proposed here, is more likely to yield well-rounded systems that are appropriately designed and implemented for their regulatory environment.

References

[1] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intell.*, vol. 1, no. 9, pp. 389–399, 2019.

[2] J. Fjeld, N. Achten, H. Hilligoss *et al.*, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI," *Research Publication No. 2020-1*, Berkman Klein Center for Internet & Society at Harvard University, 2020.

[3] C. Sanderson, D. Douglas, and Q. Lu, "Implementing responsible AI: Tensions and trade-offs between ethics aspects," in *International Joint Conference on Neural Networks*, 2023.

[4] H. Bleher and M. Braun, "Reflections on putting AI ethics into practice: How three AI ethics approaches conceptualize theory and practice," *Science and Engineering Ethics*, vol. 29, no. 3, 2023.

[5] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in AI ethics: Towards a focus on tensions," in *AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 195–200.

[6] D. Petkovic, "It is not 'accuracy vs. explainability' – we need both for trustworthy AI systems," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, pp. 46–53, 2023.

[7] E. A. Holm, "In defense of the black box," *Science*, vol. 364, no. 6435, p. 26–27, 2019.

[8] A. Barredo Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[9] C. Huang, Z. Zhang, B. Mao, and X. Yao, "An overview of artificial intelligence ethics," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 799–819, 2023.

[10] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI – two frameworks for ethical design practice," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, p. 34–47, 2020.

[11] K. Smit, M. Zoet, and J. van Meerten, "A review of AI principles in practice," in *Proc. Pacific Asia Conf. Information Systems*, 2020.

[12] V. Vakkuri, K.-K. Kemell, M. Jantunen, and P. Abrahamsson, "This is just a prototype: How ethics are ignored in software startup-like environments," *Lecture Notes in Business Information Processing*, vol. 383, pp. 195–210, 2020.

[13] R. A. Varanasi and N. Goyal, "It is currently hodgepodge: Examining AI/ML practitioners' challenges during co-production of responsible AI values," in *ACM Conf. Human Factors in Computing Systems*, 2023.

[14] S. McLennan, A. Fiske, L. A. Celi, R. Müller, J. Harder, K. Ritt *et al.*, "An embedded ethics approach for AI development," *Nature Machine Intelligence*, vol. 2, no. 9, pp. 488–490, 2020.

[15] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1083–1094, 2022.

[16] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, "The values encoded in machine learning research," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 173–184.

[17] M. W. Spratling, "Comprehensive assessment of the performance of deep learning classifiers reveals a surprising lack of robustness," *arXiv:2308.04137*, 2023.

[18] M. Jakesch, Z. Buçinca, S. Amershi, and A. Olteanu, "How different groups prioritize ethical values for responsible AI," in *ACM Conf. Fairness, Accountability, and Transparency*, 2022, pp. 310–323.

[19] W. Orr and J. L. Davis, "Attributions of ethical responsibility by artificial intelligence practitioners," *Information, Communication & Society*, vol. 23, no. 5, pp. 719–735, 2020.

[20] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, "Where Responsible AI meets reality," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 7:1–23, 2021.

[21] Australian Government (Department of Industry, Science and Resources), "Australia's Artificial Intelligence Ethics Framework," 2019, accessed 28 Nov 2023. URL: https://industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/.

[22] United Nations, "High-level advisory body on artificial intelligence," 2023, accessed 28 Nov 2023. https://www.un.org/en/ai-advisory-body.

[23] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, "Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering," *ACM Computing Surveys*, 2023.

[24] E. Bietti, "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020, pp. 210–219.

[25] D. Bruschi and N. Diomede, "A framework for assessing AI ethics with applications to cybersecurity," *AI Ethics*, vol. 3, no. 1, pp. 65–72, 2023.

[26] F. Cardinaux, C. Sanderson, and S. Bengio, "User authentication via adapted statistical models of face images," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 361–373, 2006.

[27] European Union, "Regulation (EU) 2016/679 ... (General Data Protection Regulation)," Official Journal of the European Union L 119, 4 May 2016.

[28] W. Maalej, Y. D. Pham, and L. Chazette, "Tailoring requirements engineering for responsible AI," *Computer*, vol. 56, no. 4, p. 18–27, 2023.

[29] K. Ahmad, M. Abdelrazek, C. Arora *et al.*, "Requirements engineering framework for human-centered artificial intelligence software systems," *Applied Soft Computing*, vol. 143, p. 110455, 2023.

[30] J. Horkoff, "Non-functional requirements for machine learning: Challenges and new directions," in *IEEE International Requirements Engineering Conference*, 2019.

[31] Q. Lu, L. Zhu, X. Xu, J. Whittle, and Z. Xing, "Towards a roadmap on software engineering for responsible AI," in *International Conference on AI Engineering: Software Engineering for AI*, 2022, pp. 101–112.

[32] A. Koshiyama *et al.*, "Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms," *Computer*, vol. 55, no. 4, pp. 40–50, 2022.

[33] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[34] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[35] D. Xin *et al.*, "Production machine learning pipelines: Empirical analysis and optimization opportunities," in *ACM International Conference on Management of Data*, 2021, pp. 2639–2652.

[36] M. Loi and M. Christen, "Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector," *Philosophy & Technology*, vol. 34, no. 4, pp. 967–992, 2021.

[37] G. Strang, *Introduction to Linear Algebra*. Wellesley-Cambridge, 2023.

[38] S. Buijsman, "Navigating fairness measures and trade-offs," *AI and Ethics*, 2023.

[39] T. L. Beauchamp and J. F. Childress, *Principles of Biomedical Ethics, Seventh Edition*. Oxford University Press, 2013.

[40] P. Formosa, M. Wilson, and D. Richards, "A principlist framework for cybersecurity ethics," *Computers & Security*, vol. 109, p. 102382, 2021.

[41] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.

[42] E. Seger, "In defence of principlism in AI ethics and governance," *Philosophy & Technology*, vol. 35, p. 45, 2022.

[43] M. Shea, "Principlism's balancing act: Why the principles of biomedical ethics need a theory of the good," *Journal of Medicine and Philosophy*, vol. 45, pp. 441–470, 2020.

[44] M. Karliuk, "Proportionality principles for the ethics of artificial intelligence," *AI and Ethics*, vol. 3, no. 3, pp. 985–990, 2023.

[45] J. Rawls, *A Theory of Justice*. Harvard University Press, 1971.

[46] J. Woodgate and N. Ajmeri, "Normative ethics principles for responsible AI systems: Taxonomy and future directions," *arXiv:2208.12616*, 2023.

[47] J. Flynn, "Theory and Bioethics," in *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2022.

[48] A. Blanchard, C. Thomas, and M. Taddeo, "Ethical governance of artificial intelligence for defence: normative tradeoffs for principle to practice guidance," *AI & Society*, 2024.

[49] K. Boyd, "Designing up with value-sensitive design: Building a field guide for ethical ML development," in *ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2069–2082.

[50] C. Sanderson, D. Douglas, Q. Lu, E. Schleiger, J. Whittle, J. Lacey, G. Newnham, S. Hajkowicz, C. Robinson, and D. Hansen, "AI ethics principles in practice: perspectives of designers and developers," *IEEE Trans. Technology and Society*, vol. 4, no. 2, pp. 171–187, 2023.

[51] J. Morley *et al.*, "Operationalising AI ethics: barriers, enablers and next steps," *AI & Society*, vol. 38, no. 1, pp. 411–423, 2023.

[52] C. Stix, "Actionable principles for artificial intelligence policy: Three pathways," *Science and Engineering Ethics*, vol. 27, no. 1, p. 15, 2021.