Implementing Responsible AI:
Tensions and Trade-Offs Between Ethics Aspects

Conrad Sanderson [†◇], David Douglas [‡], Qinghua Lu [†]

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

Abstract—**Many sets of ethics principles for responsible AI have been proposed to allay concerns about misuse and abuse of AI/ML systems. The underlying aspects of such sets of principles include privacy, accuracy, fairness, robustness, explainability, and transparency. However, there are potential tensions between these aspects that pose difficulties for AI/ML developers seeking to follow these principles. For example, increasing the accuracy of an AI/ML system may reduce its explainability. As part of the ongoing effort to operationalise the principles into practice, in this work we compile and discuss a catalogue of 10 notable tensions, trade-offs and other interactions between the underlying aspects. We primarily focus on two-sided interactions, drawing on support spread across a diverse literature. This catalogue can be helpful in raising awareness of the possible interactions between aspects of ethics principles, as well as facilitating well-supported judgements by the designers and developers of AI/ML systems.**

Index Terms—**AI ethics, responsible AI, interactions, tensions, trade-offs.**

I. INTRODUCTION

In order to address concerns about potentially negative impacts of the rapidly growing use of Artificial Intelligence (AI) and Machine Learning (ML) technologies throughout society [73], many governments and organisations have proposed numerous sets of AI ethics principles [16], [34]. An example is given in Fig. 1.

While such principles are well-intended, a major issue is that these principles are high-level, and do not readily provide guidance on how to concretely implement them within AI/ML systems [47], [49], [50], [65]. Furthermore, attempts at operationalising these principles reveal that many can be in tension with each other [61], [74]. This in turn can lead to suboptimal outcomes, where the designers and developers of AI/ML systems may haphazardly resolve tensions by simply selecting one principle (or its underlying aspect) to be dominant, rather than devising and incorporating well-balanced trade-offs between the principles in tension where the associated risks and benefits are gauged more thoroughly [74].

Although the available sets of AI ethics principles have notable differences in wording, a set of underlying common themes and aspects can be identified within them [21], [34]. These common aspects typically include privacy, accuracy, fairness, robustness, explainability, transparency, and accountability. Tensions can exist between such aspects which require trade-offs to be made between them; for example, increasing privacy may come at the cost of decreasing accuracy (a negative interaction). However, interactions between the aspects are not always in the forms of trade-offs: increasing robustness, for example, may increase accuracy in certain circumstances (a context-dependent interaction). Furthermore, interactions may also benefit all the involved aspects (positive interactions).

There are many observed interactions (positive, negative, and context-dependent) between various two- and three-sided combinations of the common aspects of AI ethics principles. However, descriptions of such interactions are spread across a diverse and disparate literature. This in turn may contribute to a lack of awareness among the designers and developers of AI/ML systems about the wide range and nature of the possible interactions between the underlying aspects.

In this work we aim to address the above problem in a two-part fashion. We first gather and provide definitions of the underlying aspects present in AI ethics principles, shedding light on multiple (possibly semi-compatible) definitions of each aspect. Building on [39], the various tensions, conflicts, trade-offs and other interactions between the aspects that can occur in practice are then elucidated and discussed, as supported by the available literature. Such an explicit catalogue can be

---

1. **Human, Social & Environmental Wellbeing.** AI systems should benefit individuals, society, and the environment.
2. **Human-centred Values.** AI systems should respect human rights, diversity, and the autonomy of individuals.
3. **Fairness.** AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals or groups.
4. **Privacy Protection & Security.** AI systems should respect and uphold privacy rights and regulations, and ensure the security of data.
5. **Reliability & Safety.** AI systems should reliably operate in the context of their intended purpose throughout their lifecycle.
6. **Transparency & Explainability.** There should be transparency and responsible disclosure allowing people to know when an AI system is engaging with them and/or when they are significantly impacted by an AI system.
7. **Contestability.** When an AI system significantly impacts a person, group or environment, there should be a timely process to allow people to challenge the use or output of the system.
8. **Accountability.** Human oversight of AI systems should be enabled; those responsible for the various phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the system.

Fig. 1. Summarised form of the high-level AI ethics principles proposed by the Australian Government [6].

---

helpful in raising awareness of the possible interactions, as well as facilitating well-supported judgements for implementing *responsible AI* by the designers and developers of AI/ML systems.

The structure of this paper is as follows. The descriptions of each aspect are provided in Section II. The interactions between the aspects are explored in Section III. Concluding remarks and future avenues of research are given in Section IV.

II. UNDERLYING ASPECTS

Across the many proposed sets of AI ethics principles, we focus on six underlying aspects: (a) accuracy, (b) robustness, (c) fairness, (d) privacy, (e) explainability, (f) transparency. An overview of each of these aspects is given below.

The selection of these aspects is driven by the availability of literature that covers interactions between at least two such aspects. We have elected not to explicitly cover the *accountability* aspect, due to it not being easily quantifiable, as well as its close association with the *transparency* aspect [19].

We note that while the names of the aspects may not directly match the wording used within a given set of AI ethics principles, the names can be easily mapped to the corresponding wording. For example, *robustness* can be mapped to *reliability* within the set of high-level AI ethics principles shown in Fig. 1 with a high degree of fidelity. Furthermore, based on the definitions available in the literature, we have elected to place related and/or semi-compatible definitions under one umbrella.

A. Accuracy

Given data comprising instances of samples along with associated ground-truth labels, *accuracy*, in its simplest form, can be defined as the ratio of correct classifications to the total number of classifications made by an AI/ML model. The act of classification in this case is defined as the process of mapping a given sample to one label out of a set of pre-defined labels. Accuracy is typically measured on a hold-out dataset: data that is known and available, but is not used for training the AI/ML models. This straightforward measurement approach aims to evaluate the generalisation ability of the models under known (normal) conditions [12].

In the case of binary classifications, where the possible decisions (outputs from an AI/ML model) are *true* and *false*, more complex forms of accuracy are used, such as *recall*, *precision*, *F1 score*, and *area under curve* [12], [13]. There are also task-specific measures of accuracy where the notions of *correct* and *incorrect* are less clearly defined, such as in language translation [53].

Specialisations of accuracy can also be denoted as *overall accuracy* and *group accuracy*. In the former, the entire hold-out dataset is used, where each sample is treated individually and has equal contribution to the final measurement; this is the traditional definition of accuracy [12]. In the latter, accuracy is calculated for a pre-defined group in the hold-out dataset, where the group membership is defined by one or more characteristics, such as gender and ethnicity. Within this work we use the term *accuracy* to mean *overall accuracy*, unless specified otherwise.

B. Robustness

The robustness of an AI/ML system is its ability to perform as expected under various conditions [17]. A classical definition of robustness within machine learning is the ability to maintain a high degree of accuracy under conditions that may have not been observed in the training dataset [76]. A related definition of robustness is the capacity to maintain high accuracy under a wide range of conditions [84].

A more recent definition of robustness is resistance to adversarial attacks, which may occur during training and/or usage of an AI/ML model [3], [4], [43], [67], [77]. In both cases, data is deliberately altered or constructed so that the model is induced to perform unreliably (eg. reduction in accuracy), or is biased towards producing results favoured by an attacker (eg. reduction in accuracy for a specific group). Robustness can also be defined as resiliency to label noise, where the class labels within training data are deliberately or inherently inaccurate [22], [27], [69].

The nature of the attacks can be placed into three broad categories: so-called white-box, black-box, and grey-box attacks [3]. In the white-box case, the attacks can be dependent on the type and underlying characteristics of an AI/ML model (cf. transparency in Section II-F). In the black-box case, the attacks may not have access to such details, and hence may also be probing in nature, designed to glean information about the capabilities and limitations of an AI/ML system. In the grey-box case, the attacker has partial knowledge of the system. An instructive example of adversarial attacks in the context of deep neural networks is shown in [48].

C. Fairness

Within the context of machine learning, fairness can considered as the opposite of bias, where bias can be construed as favouring a particular characteristic (eg. gender) or group. In a simplified view, two conflicting definitions of fairness exist [23], [79]: **(i)** individual fairness (eg. fairness as equality of treatment across individuals), and **(ii)** group fairness (eg. fairness as equality of outcomes across demographic groups).

Under definition (i), an AI/ML model that obtains the lowest error rates for the largest percentage of individuals is considered the fairest (ie. aims for maximising overall accuracy). Under definition (ii), an AI/ML model that obtains similar error rates across groups is the fairest (ie. aims for minimising differences between group accuracies). Representative fairness metrics include demographic parity, equality of odds, and equality of opportunity [20].

D. Privacy

Distinct from protection of sensitive data via access control means (eg. via passwords), privacy in the context of AI/ML systems can be defined as preventing inadvertent or deliberate leakage of sensitive data via algorithmic means. A straightforward approach is to remove personally identifiable information from training data (anonymisation). However, given a sufficient amount of data from external sources, it may still be possible to de-anonymise individuals [28], [51].

Approaches aiming to counteract de-anonymisation include differential privacy [31], [81], which can be roughly described as adding noise to input data, model parameters, and/or output data. Other techniques aiming towards protection of sensitive data include federated learning [42], [75] and homomorphic encryption [1], [32]. In federated learning an AI/ML model is trained using multiple decentralised computers; each computer uses only a subset of the training dataset and has no access to the other subsets. In homomorphic encryption, computations are performed on encrypted data without prior decryption of the data, thereby allowing data processing to be done by third parties (eg. cloud services). To protect data privacy in model exchanges and aggregations, federated learning can be implemented via secure multi-party computation in conjunction with homomorphic encryption [35].

E. Explainability

Explainability, in the strictest sense, can be defined as the provision of a mathematical explanation of how an AI/ML system processes given input data to produce a specific output (eg. the processing steps of a neural network). However, given the large scale and complexity of modern AI/ML pipelines (eg. deep neural networks with millions of parameters), complete explanations are very difficult to interpret by humans [2], [8]. Systems with these characteristics are often informally referred to as "black boxes" [2], [60].

Various approaches are available to provide partial explainability that allow forms of interpretability (ie. reasoning behind a given output from an AI/ML system) [2], [8]. This includes text-based and visual explanations [58], as well as post-hoc methods, such as the creation of simplified surrogate models that are easier to interpret [8], [61]. The required degree of explainability may depend on the intended audience (eg. developers, end users) as well as their background, culture, and preferences [8], [61].

F. Transparency

Transparency can be defined as the provision of partial or full details of the AI/ML system, including the type of machine learning algorithm employed, the datasets used for training and evaluation as well as the training procedure (including data pre-processing), and associated performance metrics (eg. accuracy rates on the evaluation dataset). The degree of transparency may be restricted by commercial concerns (eg. availability of source code, access to datasets) [15]. Transparency can also be defined as allowing users to be aware that they are (directly or indirectly) interacting with an AI/ML system [6].

In both definitions, transparency can be considered as a necessary component of accountability, which aims to identify the persons and/or organisations responsible for specific components of the overall AI/ML system [19].

The notion of explainability is sometimes conflated with the term transparency by designers and developers of AI/ML systems [61], where the latter term is used as a synonym for the former term (cf. Section II-E).

III. Interactions between Aspects

We describe and discuss the following interactions, based on support available in the literature:

(a) Accuracy vs. Robustness
(b) Accuracy vs. Fairness
(c) Accuracy vs. Privacy
(d) Accuracy vs. Explainability
(e) Fairness vs. Robustness
(f) Fairness vs. Privacy
(g) Fairness vs. Transparency
(h) Privacy vs. Robustness
(i) Transparency vs. Explainability
(j) Transparency vs. Privacy and Robustness

A. Accuracy vs. Robustness

Machine learning models that achieve high accuracy within the conditions defined by the training dataset (representing expected usage) can quickly degrade in performance when they do not explicitly take into account that the conditions during deployment can be significantly different due to (potentially foreseeable) changes in data acquisition circumstances [76], [84].

The accuracy of AI/ML systems can also degrade over time due to natural data drift [41]. Periodic recalibration and/or retraining of AI/ML systems may be necessary to ensure the system is working accurately in changed conditions [41], [61]. Such an approach is also known as *algorithmic maintenance* [39]. However, blindly updating training data without thoroughly testing the retrained system may result in instabilities and/or unintended changes in operation [63].

Deliberate adversarial attacks during usage of an AI/ML model may use data that is designed to fall outside of the normal conditions that the model was originally trained to handle [3], [43], [44], [77]. Increasing robustness against adversarial attacks may involve augmenting the training dataset (eg. with perturbed versions of the original data), resulting in a larger training dataset. Related approaches include the generation of additional training samples via convex linear combinations of existing samples, and taking into account during model training the immediate neighbourhood of each training sample in addition to the sample itself [3], [43], [77].

When robustness is defined as resiliency to label noise, a positive interaction between accuracy and robustness is likely to be present [27], [69]. When robustness is defined as resistance to adversarial attacks, the link between accuracy and robustness appears to have two facets. For small training datasets, taking into account robustness may increase accuracy; in contrast, for larger training datasets, taking into account robustness is likely to decrease accuracy [40], [66], [68], [83]. For small amounts of training data, the resultant models are likely to be overfitting [12], and hence incorporating robustness can act as a form of regularisation. With larger amounts of training data, the resultant models are already well-fitting and do not require further regularisation. The reduction in accuracy due to adversarial attacks can be somewhat controlled using specifically crafted ML models [57].

B. Accuracy vs. Fairness

Recent literature indicates that incorporating fairness provisions (under all definitions of fairness) never improves accuracy, and typically hinders it [11], [24], [31], [72]. Overall, the interaction can be summarised as negative. For example, with fairness defined as group fairness, as fairness increases, overall accuracy is expected to decrease [31].

It has also been observed that accuracy can be group-dependent even if the training dataset is balanced (ie. similar number of training samples for each group). Depending on the nature of a given group, accuracy for that group may be inherently lower than for other groups [11]. A notable example of this disparity is the relatively low accuracy of face recognition for females with darker skin complexion [14].

In scenarios involving medical applications, developers of an AI/ML system may need to restrict its applicability to a specific ethnic group, due to limitations in the availability of sufficient high-quality training data. Without such restriction, the system is likely to provide inaccurate outputs [61].

C. Accuracy vs. Privacy

Accuracy is typically decreased when differential privacy is employed, with the decrease being typically minor [33], [56], [64]. However, the decrease in accuracy can be considerably greater for under-represented groups (eg. minorities) [7], which in turn can affect group fairness.

Accuracy can be decreased when privacy considerations drive the minimisation of data acquisition, in order to reduce the amount of personally identifiable information that could be potentially exposed [61]. The exposure can be due to security lapses or adversarial attacks.

In cases where the use of federated learning is appropriate for addressing privacy concerns, accuracy may decrease if there are significant differences in the nature of the separate datasets held in the separate computing nodes [42], [78]. However, when AI/ML model training time is limited, the parallelised nature of federated learning can lead to increased accuracy. In contrast to using a single computing node, federated learning is quicker as multiple computing nodes are employed, thereby processing more training data within an allotted time budget [61]. Larger amounts of training data typically lead to more accurate AI/ML models [12].

D. Accuracy vs. Explainability

In cases where explainability is required (for internal development purposes and/or to provide reasoning for end users), designers and developers of AI/ML systems may choose to employ methods/models that are easier to interpret but are less accurate, rather than more accurate approaches where it is more difficult to convey the reasoning behind their outputs [60], [61]. However, accuracy and explainability are not necessarily mutually exclusive, as it is possible to devise a trade-off between them by varying model complexity [55].

Simplified (and hence more readily interpretable) models can be also used purely for the purpose of driving explanations (ie. surrogate models), rather than as replacements of the original (non-interpretable) models [9]. In this case the explanations obtained via surrogate models do not affect the accuracy of the original model. However, there is a question on the fidelity of the surrogate model, in the sense that the obtained explanations may not reflect the underlying processing within the original model [2].

E. Fairness vs. Robustness

To achieve robust functionality, the range of allowable conditions may need to be constrained, which in turn may be in tension with fairness. Using an AI/ML system beyond what the training dataset covers (eg. people with specific ethnicity) may result in unreliable operation [61].

When explicitly taking into account robustness against adversarial attacks, group fairness can be reduced, even when the training dataset is balanced across the classes. Furthermore, inherent differences in accuracy across groups can be exacerbated [11].

Within AI/ML approaches focused on text classification, robustness to targeted word substitutions can help with improving both individual fairness [82] and group fairness [25]. The targeted substitutions typically relate to words associated with a protected subgroup. Examples include robustness to changes in descriptions of sentence toxicity [25] and changes in gender [82].

F. Fairness vs. Privacy

Fairness tends to be reduced as a byproduct of employing differential privacy, as exemplified in [7], [18], [31]. Furthermore, non-fairness can be exacerbated if the original AI/ML model produces non-fair outcomes, with the decrease in fairness more pronounced for minority groups with inadequate representation [7]. However, it is possible to ameliorate this effect and achieve approximate fairness under the constraints of differential privacy with specifically constructed machine learning models [18].

G. Fairness vs. Transparency

Through increasing transparency of an AI/ML system it may become apparent to the designers or the end users that the system is unfair [8], thereby possibly leading to internal or external pressure to increase fairness. Furthermore, if the designers of an AI/ML system are required to demonstrate that the system is fair, an effective means for doing so is to increase transparency (and possibly explainability).

The appropriate definition of fairness for an AI/ML system will depend on the context of its use [70]. Disclosing how designers have understood and implemented fairness will assist the end users of an AI/ML system in evaluating whether it is likely to produce fair outcomes. This is particularly important where end users are accountable for the decisions made using an AI/ML system [71].

A possible tension between fairness and transparency is the potential for malicious or mischievous actors to exploit an AI/ML system to their own advantage based on the information disclosed about the system [37], [71]. In such cases, transparency may need to be limited to protect the fairness of the system. The designers should consider to whom the AI/ML

system should be transparent, and what information should be provided [19]. A similar tension is created if aspects of an AI/ML system are protected as trade secrets [15]. To maintain accountability where public transparency is impossible, the AI/ML system may be made transparent to a trusted auditor for review [54] (cf. Section III-J).

H. Privacy vs. Robustness

At a high level, privacy can be linked with robustness through the overarching aim of preventing the misuse of AI/ML systems. This is reflected in the literature, where both differential privacy and robustness to adversarial attacks can be incorporated into the same machine learning approach [56]. Furthermore, the link between privacy and robustness can be exploited to create a foundation for building AI/ML methods that explicitly takes into account both requirements at the same time [38].

I. Transparency vs. Explainability

Explainability can help with transparency, in the sense that having a documented understanding of the AI/ML pipeline (necessary for explainability) can be co-opted to provide abridged details (suitable for disclosure) about the AI/ML system [19]. Explainability should be understood as being targeted towards a particular audience [8]. These audiences will depend on the context within which an AI/ML system is used. Example audiences may be the end users of an AI/ML system, those affected by decisions made by an AI/ML system, or regulators. The details of an AI/ML system should be presented in formats that are easily understood and used by their target audiences [19].

J. Transparency vs. Privacy and Robustness

Increasing transparency can negatively affect both robustness and privacy, in the sense that revealing details about an AI/ML system, such as its neural network architecture and training procedure (including associated datasets), can facilitate the potential re-identification (de-anonymisation) of individuals [51], [59], as well as to perform targeted adversarial attacks in order to cause malfunction.

On the surface it may appear that opaque systems are more difficult to analyse and hence more difficult to exploit [52]. However, this view has parallels with the traditional security approach in proprietary (closed-source) software systems, which has been pejoratively referred to as *security through obscurity* [5], [46] and its efficacy questioned in contrast to open-source software [29], [45].

Given these opposing viewpoints, it may be preferable to seek a balance between the two extremes of transparency (ie. full transparency vs. no transparency), such as providing partial transparency to trusted parties, with the degree of transparency dependent on the degree of trust [54], [80].

Transparency can be also hampered through the use of proprietary datasets for training AI/ML systems, where the owner of the dataset may prevent disclosure that the dataset is used or prevent sharing it with third parties, due to commercial sensitivities and/or privacy considerations [30], [61].

IV. CONCLUDING REMARKS

Various sets of AI ethics principles typically have a common set of underlying aspects, which includes accuracy, fairness, robustness, privacy, explainability, and transparency. These aspects may interact with each other, and require designers and developers to make trade-offs between them. While some tensions between these aspects may be generally known (such as the trade-off between accuracy and privacy), the wider gamut of observed interactions between the aspects is spread across a diverse literature and not well-known as a whole.

In order to facilitate the operationalisation of AI ethics principles, in this work we have compiled and elucidated a catalogue of 10 notable interactions between the underlying aspects, primarily focusing on two-sided interactions for which there is support in the literature. While we have also covered a three-sided interaction (transparency vs. privacy and robustness), future avenues of research include the exploration of other possible three-sided interactions, such as accuracy vs. robustness vs. fairness [10], [11], and accuracy vs. privacy vs. fairness [26], [36].

Recognising the tensions and other interactions between common aspects of AI ethics principles is an important step towards operationalising these principles. The catalogue presented here can be helpful in raising awareness of the possible interactions between these aspects, as well as facilitating well-supported judgements by the designers and developers of AI/ML systems in addressing them. In follow-up work we examine several notable approaches for resolving the tensions [62].

REFERENCES

[1] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti. A survey on homomorphic encryption schemes. *ACM Computing Surveys*, 51(4):1–35, 2018.

[2] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[3] N. Akhtar, A. Mian, N. Kardan, and M. Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021.

[4] B. Alshemali and J. Kalita. Improving the reliability of deep neural networks in NLP: A review. *Knowledge-Based Systems*, 191:105210, 2020.

[5] R. Anderson. Why information security is hard – an economic perspective. In *Computer Security Applications Conference*, pages 358–365, 2001.

[6] Australian Government. Australia's artificial intelligence ethics framework, 2019. Online (accessed 10 Apr 2023). https://industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework.

[7] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. In *Neural Information Processing Systems (NeurIPS)*, pages 15479–15488, 2019.

[8] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[9] O. Bastani, C. Kim, and H. Bastani. Interpretability via model extraction. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[10] P. Benz, C. Zhang, S. Ham, A. Karjauv, G. Cho, and I. S. Kweon. Trade-off between accuracy, robustness, and fairness of deep classifiers. In *Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges*, 2021.

[11] P. Benz, C. Zhang, A. Karjauv, and I. S. Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. *Proceedings of Machine Learning Research*, 148:325–342, 2021.

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

[13] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[14] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability and Transparency (FAccT)*, pages 77–91, 2018.

[15] J. Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):1–12, 2016.

[16] M. Christoforaki and O. Beyan. AI ethics – a bird's eye view. *Applied Sciences*, 12(9):4130, 2022.

[17] A. F. Cooper, E. Moss, B. Laufer, and H. Nissenbaum. Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 864–876, 2022.

[18] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. On the compatibility of privacy and fairness. In *Fairness in User Modeling, Adaptation and Personalization (FairUMAP) Workshop*, pages 309–315, 2019.

[19] N. Diakopoulos. Transparency. In M. D. Dubber, F. Pasquale, and S. Das, editors, *The Oxford Handbook of Ethics of AI*, pages 197–213. Oxford University Press, 2020.

[20] M. Du, F. Yang, N. Zou, and X. Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2021.

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

[22] B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[23] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021.

[24] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 329–338, 2019.

[25] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

[26] X. Gu, Z. Tianqing, J. Li, T. Zhang, W. Ren, and K.-K. R. Choo. Privacy, accuracy, and model fairness trade-offs in federated learning. *Computers & Security*, 122:102907, 2022.

[27] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[28] J. Henriksen-Bulmer and S. Jeary. Re-identification attacks – a systematic literature review. *International Journal of Information Management*, 36(6):1184–1192, 2016.

[29] J.-H. Hoepman and B. Jacobs. Increased security through open source. *Communications of the ACM*, 50(1):79–83, 2007.

[30] D. Innerarity. Making the black box society transparent. *AI & Society*, 36(3):975–981, 2021.

[31] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman. Differentially private fair learning. In *International Conference on Machine Learning (ICML)*, 2019.

[32] N. Jain, K. Nandakumar, N. Ratha, S. Pankanti, and U. Kumar. PPDL - privacy preserving deep learning using homomorphic encryption. In *Joint International Conference on Data Science & Management of Data*, pages 318–319, 2022.

[33] B. Jayaraman and D. Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, pages 1895–1912, 2019.

[34] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[35] M. Keller and K. Sun. Secure quantized training for deep learning. *Proceedings of the 39th International Conference on Machine Learning, PMLR*, 162:10912–10938, 2022.

[36] A. Kesari. The privacy-fairness-accuracy frontier: A computational law & economics toolkit for making algorithmic tradeoffs. In *ACM Symposium on Computer Science and Law*, pages 77–85, 2022.

[37] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165, 2017.

[38] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*, pages 656–673, 2019.

[39] J. A. Lee. *Artificial Neural Networks In Public Policy: Towards An Analytical Framework*. PhD thesis, George Mason University, 2020.

[40] S. Lei, H. Zhang, K. Wang, and Z. Su. How training data affect the accuracy and robustness of neural networks for image classification. *OpenReview pre-print (https://openreview.net/forum?id=HklKWhC5F7)*, 2019.

[41] G. A. Lewis, S. Echeverría, L. Pons, and J. Chrabaszcz. Augur: A step towards realistic drift detection in production ML systems. In *IEEE/ACM International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*, pages 37–144, 2022.

[42] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[43] Y. Li, M. Cheng, C.-J. Hsieh, and T. C. M. Lee. A review of adversarial attack and defense for classification methods. *The American Statistician*, 76(4):329–345, 2022.

[44] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[45] R. T. Mercuri. Trusting in transparency. *Communications of the ACM*, 48(5):15–19, 2005.

[46] R. T. Mercuri and P. G. Neumann. Security by obscurity. *Communications of the ACM*, 46(11):160, 2003.

[47] P. Mikalef, K. Conboy, J. E. Lundström, and A. Popoviĉ. Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, 31(3):257–268, 2022.

[48] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[49] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal. 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, 2020.

[50] J. Morley, L. Kinsey, A. Elhalal, F. Garcia, M. Ziosi, and L. Floridi. Operationalising AI ethics: barriers, enablers and next steps. *AI & Society*, 38(1):411–423, 2023.

[51] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[52] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *ACM Asia Conference on Computer and Communications Security*, pages 506–519, 2017.

[53] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[54] F. Pasquale. *The Black Box Society*. Harvard University Press, 2015.

[55] D. Petkovic. It is not 'accuracy vs. explainability' – we need both for trustworthy AI systems. *IEEE Transactions on Technology and Society*, 4(1):46–53, 2023.

[56] H. Phan, M. T. Thai, H. Hu, R. Jin, T. Sun, and D. Dou. Scalable differential privacy with certified robustness in adversarial learning. In *International Conference on Machine Learning (ICML)*, pages 7683–7694, 2020.

[57] R. Pinot, L. Meunier, F. Yger, C. Gouy-Pailler, Y. Chevaleyre, and J. Atif. On the robustness of randomized classifiers to adversarial examples. *Machine Learning*, 111(9):3425–3457, 2022.

[58] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *International Conf. Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, 2016.

[59] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1):1–9, 2019.

[60] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[61] 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 Transactions on Technology and Society*, 4(2):171–187, 2023. DOI: 10.1109/TTS.2023.3257303.

[62] C. Sanderson, E. Schleiger, D. Douglas, P. Kuhnert, and Q. Lu. Resolving ethics trade-offs in implementing responsible AI. *arXiv:2401.08103*, 2024.

[63] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *Neural Information Processing Systems (NeurIPS)*, pages 2503–2511, 2015.

[64] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[65] K. Smit, M. Zoet, and J. van Meerten. A review of AI principles in practice. In *Proc. Pacific Asia Conference on Information Systems*, 2020.

[66] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is robustness the cost of accuracy? — a comprehensive study on the robustness of 18 deep image classification models. *Lecture Notes in Computer Science*, 1216:644–661, 2018.

[67] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[68] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

[69] A. Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Neural Information Processing Systems (NeurIPS)*, pages 5596–5605, 2017.

[70] M. Veale and R. Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):1–17, 2017.

[71] M. Veale, M. Van Kleek, and R. Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[72] C. Wadsworth, F. Vera, and C. Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. In *Fairness, Accountability, and Transparency in Machine Learning*, 2018.

[73] J. Whittle. Is your software valueless? *IEEE Software*, 36(3):112–115, 2019.

[74] 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*, pages 195–200, 2019.

[75] A. Wilson, M. Solh, and M. Moh. Exploration of privacy preserving deep learning framework for computer vision tasks. In *ACM Southeast Conference*, pages 130–137, 2022.

[76] Y. Wong, M. T. Harandi, C. Sanderson, and B. C. Lovell. On robust biometric identity verification via sparse encoding of faces: Holistic vs local approaches. In *International Joint Conference on Neural Networks (IJCNN)*, 2012.

[77] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *Int. Journal of Automation and Computing*, 17(2):151–178, 2020.

[78] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning. *ACM Trans. Intelligent Systems and Technology*, 10(2):1–19, 2019.

[79] S. Yeom and M. C. Tschantz. Avoiding disparity amplification under different worldviews. In *ACM Conference on Fairness, Accountability, and Transparency (FAacT)*, pages 273–283, 2021.

[80] M. Young, L. Rodriguez, E. Keller, F. Sun, B. Sa, J. Whittington, and B. Howe. Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 191–200, 2019.

[81] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex. Differentially private model publishing for deep learning. In *IEEE Symposium on Security and Privacy*, pages 332–349, 2019.

[82] M. Yurochkin, A. Bower, and Y. Sun. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations (ICLR)*, 2020.

[83] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

[84] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller. Deep learning for environmentally robust speech recognition. *ACM Trans. Intelligent Systems and Technology*, 9(5):1–28, 2018.